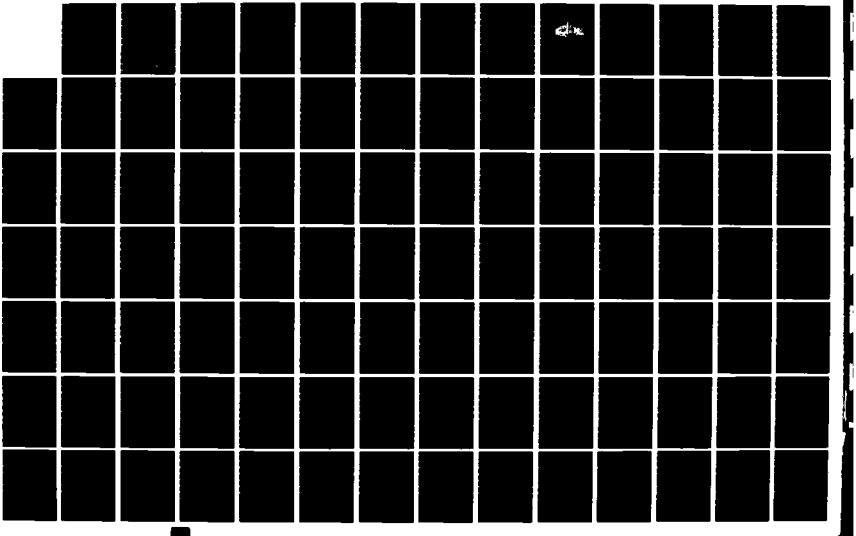
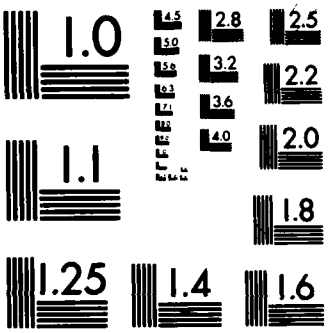
AD-A130 770 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM (1982) 1/11
RESEARCH REPORTS VOLUME. (U) SOUTHEASTERN CENTER FOR
ELECTRICAL ENGINEERING EDUCATION INC S.
UNCLASSIFIED W D PEELE ET AL. OCT 82 AFOSR-TR-83-0614 F/G 5/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

DTIC FILE COPY
AD A130770

(1)

1982 UNITED STATES AIR FORCE SUMMER FACULTY RESEARCH PROGRAM

(U) REPORTS VOLUME 2 OF 2

SPONSORED BY
AIR FORCE SYSTEMS COMMAND
AIR FORCE ENGINEERING AND SERVICES CENTER
(AFSC)

CONDUCTED BY
SOUTHEASTERN CENTER FOR
ELECTRICAL ENGINEERING EDUCATION
(SCEEE)

DTIC
JUL 27 1983
A

EDITED BY
EARL T. GLOVER
SCEEE DIRECTORS

EDITED BY
MAJOR GENERAL R. BOND
AFSC/DE

This document has been approved for public release and sale; its distribution is unlimited.

PRESS

83 07 26 141

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 83-0614	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) UNITED STATES AIR FORCE SUMMER FACULTY RESEARCH PROGRAM RESEARCH REPORT, VOLUME II		5. TYPE OF REPORT & PERIOD COVERED FINAL 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Warren D. Peele Earl L. Steele		8. CONTRACT OR GRANT NUMBER(s) F49620-82-C-0035
9. PERFORMING ORGANIZATION NAME AND ADDRESS Southeastern Center for Electrical Engineering Education 11th and Massachusetts Ave St. Cloud, Florida 32769		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2301/D5
11. CONTROLLING OFFICE NAME AND ADDRESS AFOSR/XOT Building 410 Bolling AFB, DC 20332		12. REPORT DATE October 1982 13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

DTIC
ELECTE
JUL 27 1983
S D
A

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The United States Air Force Summer Faculty Research Program (USAF-SFRP) is s program designed to introduce university, college, and technical institute faculty members to Air Force research. This is accomplished by faculty competittion on a nationally advertised competitive basis for a ten-week assignment during the summer intersession to perform research at Air Force laboratories/centers. Each assignment is in a subject area and at an Air Force facility mutually agreed upon by the faculty member and the Air Force. In addition to compensation and travel expenses, a cost of living

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE UNCLASSIFIED (over)

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. (cont)

allowance is also paid. The USAF-SFRP is sponsored by the Air Force Office of Scientific Research/Air Force Systems Command, United States Air Force, and is conducted by the Southeastern Center for Electrical Engineering Education(SCEEE).

AFOSR-TR- 83-0614

Approved for public release;
distribution unlimited.

AIR FO[illegible]
[illegible]
[illegible]
[illegible]
Dist[illegible]
MATTHEW J.
Chief, Technical Information Division

1982 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM

Conducted by

Southeastern Center for

Electrical Engineering Education

under

USAF Contract Number F49620-82-C-0035

RESEARCH REPORTS

Volume II of II

Submitted to

Air Force Office of Scientific Research

Bolling Air Force Base

Washington D.C.

by

Southeastern Center for
Electrical Engineering Education

October 1982

DTIC
COPY
INSPECTED
2

DTIC
ELECTE
JUL 2 7 1983
A

Accession
NTIS
DTIC
Unannounced
Justification
By
Distribution/
Availability Codes
Dist
A-1

PREFACE

The United States Air Force Summer Faculty Research Program (USAF-SFRP) is a program designed to introduce university, college, and technical institute faculty members to Air Force research. This is accomplished by faculty competition on a nationally advertised competitive basis for a ten-week assignment during the summer intersession to perform research at Air Force laboratories/centers. Each assignment is in a subject area and at an Air Force facility mutually agreed upon by the faculty member and the Air Force. In addition to compensation and travel expenses, a cost of living allowance is also paid. The USAF-SFRP is sponsored by the Air Force Office of Scientific Research/Air Force Systems Command, United States Air Force, and is conducted by the Southeastern Center for Electrical Engineering Education (SCEEE).

The specific objectives of the 1982 USAF-SFRP are:

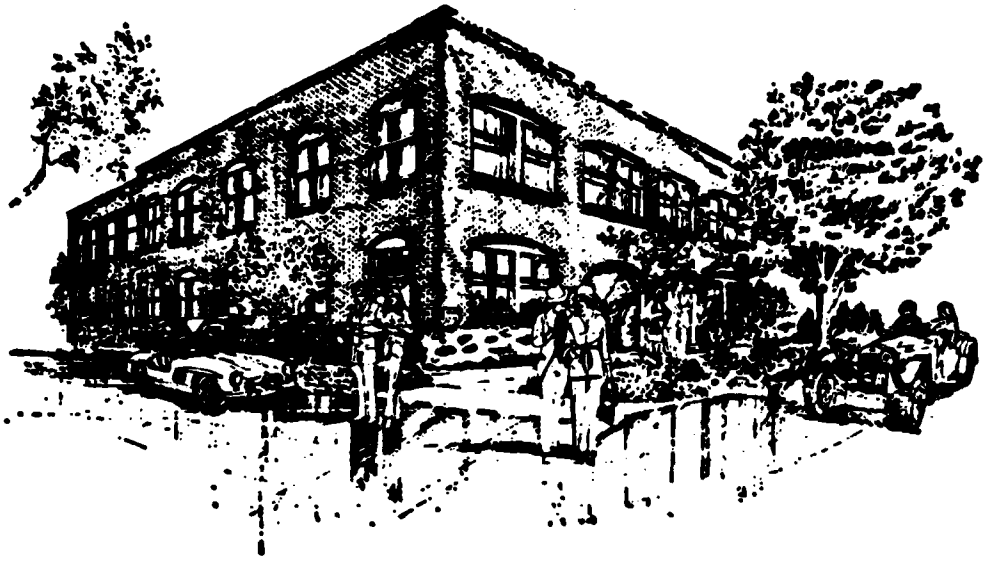
(1) To develop the basis for continuing research of interest to the Air Force at the faculty member's institution.

(2) To further the research objectives of the Air Force.

(3) To stimulate continuing relations among faculty members and their professional peers in the Air Force.

(4) To enhance the research interests and capabilities of scientific and engineering educators.

In the 1979 summer program, 70 faculty members participated, and in the 1980 and 1981 programs, 87 faculty members participated. In 1982, 91 faculty and 17 students participated. These researchers were assigned to 25 USAF laboratories/centers across the country. This two volume document is a compilation of the final reports written by the assigned faculty members about their summer research efforts.



SCEEE
©
1981

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Prof. Richard G. Absher Professor University of Vermont Electrical Engineering Department Burlington, VT 05405 (802) 658-3330	Degree: Ph.D., Electrical Engineering, 1967 Specialty: Digital Systems, Modern Control Theory, Biomedical Engineering Assigned: RADC/Griffiss
Dr. Milton J. Alexander Professor Auburn University Management Department Auburn, AL 36830 (205) 826-4730	Degree: D.B.A., Management, 1968 Specialty: Management Information Systems, Operational Research Assigned: LMC
Dr. Gary L. Allen Assistant Professor Old Dominion University Psychology Department Norfolk, VA 23508 (804) 440-4444	Degree: Ph.D., Psychology, 1979 Specialty: Cognitive Psychology, Perception, Developmental Psychology Assigned: HRL/B
Dr. Silverio P. Almeida Professor Virginia Tech. Physics Department Blacksburg, VA 24061 (703) 961-5473	Degree: Ph.D., Physics Specialty: Laser Applications to Biophysics and Pattern Recognitions Problems Assigned: SAM
Dr. Muluneh Azene Assistant Professor Southern University Civil Engineering Department Baton Rouge, LA 70813 (504) 771-5871	Degree: Ph.D., Structural Mechanics, 1979 Specialty: Finite Element Analysis Assigned: ESC
Dr. Francesco L. Bacchialoni Associate Professor University of Lowell Electrical Engineering Dept. Lowell, MA 01854 (617) 452-5000	Degree: Ph.D., Engineering, 1946 Specialty: Control Systems, Signal Processing Assigned: GL
Dr. A. Terry Bahill Associate Professor Carnegie-Mellon University Electrical Engineering Dept. Pittsburg, PA 15213 (412) 578-2536	Degree: Ph.D., Electrical Engineering, 1975 Specialty: Biological Control Systems Assigned: AMRL

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Dr. Mason G. Bailey Associate Professor Western Carolina University Accounting/Information Systems Dept. Cullowhee, NC 28723 (704) 227-7401	Degree: Ph.D., Computer Science, 1978 Specialty: Computer Science, Operations Research Assigned: HRL/WP
Dr. Pradip M. Bakshi Professor Boston College Physics Department Chestnut Hill, MA 02167 (617) 969-0100	Degree: Ph.D., Theoretical Physics, 1962 Specialty: Theoretical Plasma Physics Assigned: GL
Dr. Gust Bambakidis Assistant Professor Wright State University Physics Department Dayton, OH 45435 (513) 873-2954	Degree: Ph.D., Physics, 1966 Specialty: Theoretical Physics Assigned: ML
Prof. Albert W. Biggs Professor University of Kansas Electrical Engineering Department Lawrence, KS 66045 (913) 864-4615	Degree: Ph.D., Electrical Engineering, 1965 Specialty: Microwaves, Antennas, and Wave Propagation Assigned: RADC/Hanscom
Dr. Jack J. Bourquin Associate Professor University of Missouri Electrical Engineering Department Rolla, MO 65401 (314) 341-4548	Degree: Ph.D., Electrical Engineering, 1968 Specialty: Circuits and Systems Assigned: AL
Dr. Willie A. Bragg Assistant Professor University of Cincinnati Early Childhood & Special Education Dept. Cincinnati, OH 45221 (513) 475-4542	Degree: Ph.D., Special Education, 1979 Specialty: Mental Retardation, Early Childhood Education Assigned: SAM
Dr. Eugene F. Brown Associate Professor VPI & SU Mechanical Engineering Department Blacksburg, VA 24061 (703) 961-7199	Degree: Ph.D., Mechanical Engineering, 1968 Specialty: Fluid Mechanics, Computational Fluid Dynamics Assigned: APL

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Dr. Sherman D. Brown Professor University of Illinois Ceramic Engineering Department Urbana, IL 61801 (217) 333-4766	Degree: Ph.D., Chemical Engineering, 1957 Specialty: Brittle Fracture, Ceramic Coatings, Glasses, High Temperature Materials Assigned: ML
Dr. Louis W. Buckalew Assistant Professor Alabama A&M University Psychology Department Normal, AL 35762 (205) 859-7451	Degree: M.S. & (ABD), Physiological Psychology, 1969 & (1979) Specialty: Physiological Psychology, Psychopharm, Human Engineering Assigned: AMRL
Dr. Donald L. Byrkett Associate Professor Miami University Systems Analysis Department Oxford, OH 45046 (513) 523-2325	Degree: Ph.D., Industrial Engineering, 1974 Specialty: Mathematical Modeling, Management Science Assigned: LC
Dr. Jimmie J. Cathey Associate Professor University of Kentucky Electrical Engineering Department Lexington, KY 40506 (606) 258-4684	Degree: Ph.D., Electrical Engineering, 1972 Specialty: Analysis, Design, and Control of Electrical Machinery Assigned: APL
Dr. Roger E. Cavallo Associate Professor State University of New York Computer & Information Science Dept. Utica, NY 13502 (315) 792-3315	Degree: Ph.D., Systems Science, 1978 Specialty: Systems Methodology Assigned: RADC/G
Dr. Junho Choi Assistant Professor Florida Institute of Technology Dept. of Electrical Engineering & CP Melbourne, FL 32901 (305) 723-3701	Degree: Ph.D., Electrical Engineering, 1978 Specialty: Digital Signal Processing, Control Systems Assigned: ESMC
Dr. David L. Cleeton Assistant Professor Oberling College Economics Department Oberlin, OH 44074 (216) 775-8482	Degree: Ph.D., Economics, 1980 Specialty: Microeconomics, Labor Economics, Public Finance Assigned: BRMC

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Dr. Gregory M. Corso Assistant Professor Georgia Institute of Technology Psychology Department Atlanta, GA 30332 (404) 894-4260	Degree: Ph.D., Engineering Psychology, 1978 Specialty: Human Performance Theory, Engineering Psychology Assigned: AMRL
Dr. Robert A. Douglas Professor North Carolina State University Civil Engineering Department Raleigh, NC 27650 (919) 737-2331	Degree: Ph.D., Engineering, 1956 Specialty: Mechanics of Solids, Impact and Stress Wave Propagation Assigned: ESC
Dr. Hamed K. Eldin Professor Oklahoma State University Industrial Engineering & Management Dept. Stillwater, OK 74078 (405) 624-6055	Degree: Ph.D., Industrial Engineering, 1951 Specialty: Project Management, Economic Analysis Inventory Theory Assigned: BRMC
Dr. John D. Enderle Assistant Professor North Dakota State University Electrical Engineering Department Fargo, ND 58105 (701) 237-7689	Degree: Ph.D., Biomedical Engineering, 1980 Specialty: Epidemiology, Signal Processing, Statistical Methods & Modeling Assigned: SAM
Dr. Fernando E. Fagundo Assistant Professor University of Florida Civil Engineering Department Gainesville, FL 32611 (904) 392-0851	Degree: Ph.D., Structural Engineering, 1980 Specialty: Analysis and Design of Reinforced Concrete Structural Systems, Assigned: ESC
Dr. Hubert S. Feild Associate Professor Auburn University Management Department Auburn, AL 36830 (205) 826-4039	Degree: Ph.D., Psychology, 1973 Specialty: Industrial Psychology Assigned: LMDC
Dr. Mack Felton, Jr. Chairman Southern University Biology Department New Orleans, LA 70126 (504) 282-4401	Degree: Ph.D., Microbiology, 1973 Specialty: Microbiology Assigned: SAM

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Prof. Eugene H. Galluscio Professor Northwest Missouri State University Psychology Department Maryville, MO 64468 (816) 582-7141	Degree: Ph.D., Physiological Psychology,1970 Specialty: Neuropsychology, Brain Function, Visual Information Processing Assigned: HRL/Williams
Dr. Ronald L. Greene Assistant Professor University of New Orleans Physics Department New Orleans, LA 70148 (504) 286-6714	Degree: Ph.D., Physics, 1974 Specialty: Plasma Spectroscopy Assigned: AL
Dr. Keith C. Hanson Professor Lamar University Chemistry Department Beaumont, TX 77710 (713) 838-8267	Degree: Ph.D., Chemistry, 1967 Specialty: Organic Chemistry, Measurement of Physical Properties Assigned: FJSRL
Dr. Donald J. Healy Assistant Professor Georgia Institute of Technology Electrical Engineering Department Atlanta, GA 30332 (404) 894-2923	Degree: Ph.D., Electrical Engineering, 1981 Specialty: Image Processing (Communications) Assigned: AD
Prof. Ervin Hindin Professor Washington State University Civil & Environmental Engineering Dept. Pullman, WA 99164 (509) 335-7028	Degree: M.S., Environmental Chemistry, 1956 Specialty: Aquatic Chemistry, Water Supply Systems Assigned: ESC
Dr. Manuel A. Huerta Associate Professor University of Miami Physics Department Coral Gables, FL 33124 (305) 284-2323	Degree: Ph.D., Physics, 1970 Specialty: Plasma Physics, Fluid Mechanics, Electromagnetic Theory Assigned: AD
Dr. Francis J. Jankowski Professor Wright State University Engineering Department Dayton, OH 45435 (513) 873-2097	Degree: Sc.D., Physics, 1949 Specialty: Systems Engineering, Nuclear Engineering. Mechanical Engineering, Human Factors Engineering Assigned: WL

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Dr. Eric R. Johnson Assistant Professor Virginia Tech. Aerospace & Ocean Engineering Dept. Blacksburg, VA 24061 (703) 961-6699	Degree: Ph.D., Applied Mechanics, 1976 Specialty: Structural Stability & Composite Structures Assigned: FDL
Dr. James J. Kane Associate Professor Wright State University Chemistry Department Dayton, OH 45435 (513) 873-2352	Degree: Ph.D., Organic Chemistry, 1960 Specialty: Organic Chemistry Polymer Chemistry Assigned: ML
Dr. Samuel G. Kelly Assistant Professor University of Notre Dame Aerospace & Mechanical Engineering Dept. Notre Dame, IN 46556 (219) 239-7678	Degree: Ph.D., Engineering Mechanics, 1979 Specialty: Nonlinear Mechanics Assigned: FDL
Dr. Dennis M. Kern Assistant Professor University of Georgia Statistics & Computer Science Dept. Athens, GA 30602 (404) 542-2911	Degree: Ph.D., Statistics, 1976 Specialty: Microprocessors Assigned: SAM
Dr. Stuart T. Klapp Professor California State University Psychology Department Hayward, CA 94542 (415) 881-3484	Degree: Ph.D., Experimental Psychology, 1969 Specialty: Cognition and Human Performance Assigned: HRL/WP
Dr. Jerome Knopp Assistant Professor University of Missouri-Rolla Electrical Engineering Department Rolla, MO 65401 (314) 341-45439	Degree: Ph.D., Electrical Engineering Specialty: Electro-Optics Assigned: RADC/Griffiss
Dr. Keith Koenig Assistant Professor Mississippi State University Aerospace Engineering Department Mississippi State, MS 39762 (601) 323-3623	Degree: Ph.D., Aeronautics, 1978 Specialty: Laser Velocimetry, Separated Flow Assigned: AD

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Dr. Letitia J. Korbly Assistant Professor University of Alabama Mathematics Department Birmingham, AL 35294 (205) 934-2154	Degree: Ph.D., Mathematics, 1976 Specialty: Applied Math, Partial Differential Equations Assigned: WL
Dr. Philip Langer Associate Professor University of Colorado Ed. Psychology Department Boulder, CO 80309 (303) 492-8748	Degree: Ph.D., Education/Psychology, 1957 Specialty: Instructional Systems Assigned: HRL/L
Dr. Stephen F. Lin Associate Professor North Carolina Central University Chemistry Department Durham, NC 27707 (919) 683-6463	Degree: Ph.D., Physical Chemistry, 1970 Specialty: Physical Chemistry Assigned: RPL
Dr. Michael L. Lobb Assistant Professor University of Texas Graduate School of Social Work Arlington, TX 76019-0129 (817) 273-2707	Degree: Ph.D., Psychology, 1978 Specialty: Statistics and Experimental Design Assigned: SAM
Dr. D. J. Medeiros Assistant Professor Penn State University Industrial Engineering Department University Park, PA 16802 (814) 863-2364	Degree: Ph.D., Industrial Engineering, 1981 Specialty: Scheduling, Computer Simulation of Production/ Manufacturing Systems Assigned: ML
Dr. Thomas M. Miller Associate Professor University of Oklahoma Physics and Astronomy Department Norman, OK 70319 (405) 325-3961	Degree: Ph.D., Physics, 1968 Specialty: Experimental Atomic Collisions Assigned: GL
Dr. Michael J. Moloney Professor Rose-Hulman Institute of Technology Physics Department Terre Haute, IN 47803 (812) 877-1511	Degree: Ph.D., Physics, 1966 Specialty: Solid State Electronic Devices, Modern Physics Assigned: AL

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Dr. Luigi Morino Professor Boston University Aerospace & Mechanical Engineering Dept. Boston, MA 02215 (617) 353-3069	Degree: Ph.D., Aerospace Engineering, 1966, Ph.D., Mechanical Engineering, 1963 Specialty: Unsteady Aerodymanics and Structural Dynamics Assigned: AD
Dr. David S. Moroi Professor Kent State University Physics Department Kent, OH 44242 (216) 672-2596	Degree: Ph.D., Atomic Physics, 1959 Specialty: Atomic Physics, Transport Phenomena in Liquid Crystals Assigned: ML
Dr. H. Troy Nagle, Jr. Professor Auburn University Electrical Engineering Department Auburn, AL 36849 (205) 826-4330	Degree: Ph.D., Electrical Engineering, 1968 Specialty: Computer Engineering Assigned: SAM
Dr. Eugene M. Norris Associate Professor George Mason University Mathematical Science Department Fairfax, VA 22030 (703) 323-2262	Degree: Ph.D., Mathematics, 1969 Specialty: Theoretical Computer Science Assigned: AL
Dr. Thomas H. Ortmeyer Assistant Professor Clarkson College Electrical Engineering Department Potsdam, NY 13676 (315) 268-6536	Degree: Ph.D., Electrical Engineering, 1980 Specialty: Electric Machinery and Power Systems Assigned: APL
Dr. Dean H. Owen Associate Professor Ohio State University Psychology Department Columbus, OH 43210 (614) 422-7641	Degree: Ph.D., Experimental Psychology, 1965 Specialty: Perception, Psychophysics, Flight Simulation Assigned: HRL/Williams
Dr. Charles K. Parsons Assistant Professor Georgia Institute of Technology College of Management Atlanta, GA 30332 (404) 894-2619	Degree: Ph.D., Psychology, 1980 Specialty: Item Response Theory Assigned: HRL/Brooks

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Dr. Surgounda A. Patil Professor Tennessee Technical University Math & Computer Science Department Cookville, TN 38501 (615) 528-3593	Degree: Ph.D., Statistics, 1966 Specialty: Statistical Problems Applicable to Engineering and Environment Assigned: WL
Dr. Lakhpat R. Pujara Associate Professor Wilberforce University Engineering/Mathematics Department Wilberforce, OH 45384 (513) 376-2911	Degree: Ph.D., Mathematics, 1971 M.S., Systems Engineering, 1981 Specialty: Control Systems Assigned: FDL
Dr. Zahir Qureshi Assistant Professor Rust College Biology Department Holly Springs, MS 36835 (601) 252-4661	Degree: Ed.D., Biochemistry, 1975 Specialty: Biochemistry Education Assigned: AMRL
Dr. Ronald L. Remke Assistant Professor University of South Floirda Electrical Engineering Department Tampa, FL 33620 (813) 974-2581	Degree: Ph.D., Electrical Engineering, 1977 Specialty: Electronic Devices, Materials, and Processes/ Thin Films Assigned: RADC/Griffiss
Dr. Richard W. Rice Assistant Professor Clemson University Chemical Engineering Department Clemson, SC 29631 (803) 656-3055	Degree: Ph.D., Chemical Engineering, 1972 Specialty: Catalysis, Chemical Kinetics Assigned: ESC
Dr. Gerhard X. Ritter Associate Professor University of Florida Math & Computer Science Department Gainesville, FL 32611 (904) 392-4988	Degree: Ph.D., Mathematics, 1971 Specialty: Pattern Recognition, Applied Mathematics Assigned: AD
Dr. John M. Russell Assistant Professor University of Oklahoma School of Aero. Mech. and Nuclear Eng. Norman, OK 70319 (405) 325-5011	Degree: Ph.D., Fluid Mechanics, 1981 Specialty: Incompressible Flows, Boundry Layers, Stability, Turbulence Assigned: FDL

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Dr. Sarwan S. Sandhu Assistant Professor University of Dayton Chemical Engineering Department Dayton, OH 45469 (513) 229-2627	Degree: Ph.D., Combustion, 1973 Specialty: Combustion Assigned: APL
Dr. Robert E. Schlegel Assistant Professor University of Oklahoma Industrial Engineering Department Norman, OK 70319 (405) 325-3721	Degree: Industrial Engineering, 1980 Specialty: Human Factors Assigned: SAM
Dr. Eugene P. Schram Associate Professor Ohio State University Chemistry Department Columbus, OH 43210 (614) 222-1487	Degree: Ph.D. Inorganic Chemistry Specialty: Organometallic Chemistry Assigned: AL
Dr. K. Sam. Shanmugan Associate Professor University of Kansas Electrical Engineering Department Lawrence, KS 66045 (913) 864-4832	Degree: Ph.D. Specialty: Communication Systems Engineering Assigned: RADC/Griffiss
Dr. Trilochan Singh Associate Professor Wayne State University Mechanical Engineering Department Detroit, MI 48202 (313) 577-3845	Degree: Ph.D. Mechanical Engineering, 1970 Specialty: Combustion, Heat Transfer, Energy Conservation Assigned: APL
Dr. Boghos D. Sivazlian Professor University of Florida Industrial & Systems Engineering Dept. Gainesville, FL 32611 (904) 392-1464	Degree: Ph.D., Operations Research, 1966 Specialty: Operations Research, Math. Modeling, Military Problems Assigned: AD
Dr. Stanley L. Spiegel Assistant Professor University of Lowell Mathematics Department Lowell, MA 01854 (617) 738-5000	Degree: Ph.D., Physics, 1966 Specialty: Numerical Modeling and Computer Simulation of Geophysical Problems Assigned: GL

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Dr. Alexander P. Stone Professor University of New Mexico Mathematics Department Albuquerque, NM 87131 (505) 277-4643	Degree: Ph.D., Mathematics, 1965 Specialty: Differential Equations, Differential Geometry Assigned: WL
Dr. Alfred G. Striz Assistant Professor University of Oklahoma Aero Mechanical Nuclear Eng. Dept. Norman, OK 70319 (405) 325-5011	Degree: Ph.D., Aeronautics and Astronautics, 1981 Specialty: Aeroelasticity, Finite Elements, Aerospace Structure Assigned: AD
Dr. Patrick J. Sweeney Associate Professor University of Dayton Engineering Management Department Dayton, OH 45469 (513) 229-2238	Degree: Ph.D., Mechanical Engineering, 1977 Specialty: Simulation, O.R., Management Assigned: AMRL
Dr. Richard H. Tipping Professor University of Nebraska Physics Department Omaha, NE 68182 (402) 554-2510	Degree: Ph.D., Physics, 1969 Specialty: Molecular Spectroscopy, Atmospheric Physics Assigned: GL
Dr. Edward A. Walters Associate Professor University of New Mexico Chemistry Department Albuquerque, NM 87131 (505) 277-5239	Degree: Ph.D., Chemistry, 1966 Specialty: Physical Chemistry Assigned: WL
Dr. Kai Wang Associate Professor Wayne State University Mathematics Department Detroit, MI 48202 (313) 577-3193	Degree: Ph.D., Mathematics, 1972 Specialty: Spectral of graphs, Group Matrices Assigned: APL
Dr. Edward R. Ward, Jr. Associate Professor Alabama A & M University Biology Department Normal, AL 35762 (205) 859-7268	Degree: Ph.D., Microbiology, 1975 Specialty: Pathogenic Microbiology, Immunology Assigned: SAM

NAME/ADDRESS	DEGREE, SPECIALTY, LABORATORY ASSIGNED
Dr. Thomas E. Webb Professor Ohio State University Physiological Chemistry Department Columbus, OH 43210 (614) 422-0103	Degree: Ph.D., Biochemistry, 1961 Specialty: Biochemistry, Molecular Bioology, Carcinogenesis Assigned: AMRL
Prof. David Weimer Associate Professor Ohio Northern University Physics Department Ada, OH 45810 (419) 634-9921	Degree: M.S., Physics, 1946 Specialty: Gas dynamics, Shock Wave Phenomena, Electrooptical Effects and Instrumentation Assigned: APL
Dr. Thomas A. Wiggins Professor Penn State University Physics Department University Park, PA 16802 (814) 865-5233	Degree: Ph.D., Physics, 1980 Specialty: Physical Optics, Molecular Spectra Assigned: FJSRL
Dr. David H. Williams Assistant Professor University of Texas Electrical Engineering Department El Paso, TX 79968 (915) 747-5470	Degree: Ph.D., Electrical Engineeing, 1977 Specialty: Computer Graphics Assigned: WL
Dr. Robert E. Willis Assistant Professor Mercer University Physics Department Macon, GA 31207 (912) 744-2704	Degree: Ph.D., Physics, 1979 Specialty: Microwave Spectroscopy Assigned: AEDC
Dr. David C. Wilson Associate Professor University of Florida Mathematics Department Gainesville, FL 32611 (904) 392-6035	Degree: Ph.D., Mathematics, 1969 Specialty: Dimension Theory Assigned: FDL
Dr. Rama K. Yedavalli Assistant Professor Stevens Institute of Technology Mechanical Engineering Department Hoboken, NJ 07030 (201) 420-5574	Degree: Ph.D., Aerospace Engineering, 1981 Specialty: Sensitivity Theory for Linear Multivariable Optimal Control Systems Assigned: FDL

PARTICIPANT LABORATORY ASSIGNMENT

1982 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM

AERO PROPULSION LABORATORY
(Wright-Patterson Air Force Base)

1. Dr. Eugene Brown - Virginia Polytechnic Institute & State University
2. Dr. Jimmie Cathey - University of Kentucky
3. Dr. Dennis Flentge - Cedarville College
4. Dr. Thomas Ortmeyer - Clarkson College
5. Dr. Sarwan Sandhu - University of Dayton
6. Dr. Kai Wang - Wayne State University
7. Prof. David Weimer - Ohio Northern University

AEROSPACE MEDICAL RESEARCH LABORATORY
(Wright-Patterson Air Force Base)

1. Dr. Terry Bahill - Carnegie-Mellon University
2. Dr. Louis Buckalew - Alabama A&M University
3. Dr. Gregory Corso - Georgia Institute of Technology
4. Dr. Andris Freivalds - Pennsylvania State University
5. Dr. Zahir Quershi - Rust College
6. Dr. Patrick Sweeney - University of Dayton
7. Dr. Thomas Webb - Ohio State University

ARMAMENT DIVISION
(Eglin Air Force Base)

1. Dr. Donald Healy - Georgia Institute of Technology
2. Dr. Manuel Huerta - University of Miami
3. Dr. Keith Koenig - Mississippi State University
4. Dr. Luigi Morino - Boston University
5. Dr. Gerhard Ritter - University of Florida
6. Dr. Boghos Sivazlian - University of Florida
7. Dr. Alfred Striz - University of Oklahoma

ARNOLD ENGINEERING DEVELOPMENT CENTER
(Arnold Air Force Station)

1. Dr. Larry Forney - Georgia Institute of Technology
2. Dr. Robert Willis - Mercer University

AVIONICS LABORATORY
(Wright-Patterson Air Force Base)

1. Dr. Jack Bourquin - University of Missouri
2. Dr. Albert Frasca - Wittenberg University
3. Dr. Ronald Greene - University of New Orleans
4. Dr. Michael Moloney - Rose-Hulman Institute of Technology
5. Dr. Eugene Norris - George Mason University
6. Dr. Eugene Schram - Ohio State University

BUSINESS RESEARCH MANAGEMENT CENTER
(Wright-Patterson Air Force Base)

1. Dr. David Cleeton - Oberlin College
2. Dr. Hamed Eldin - Oklahoma State University

PARTICIPANT LABORATORY ASSIGNMENT (Continued)

EASTERN SPACE & MISSILE CENTER
(Patrick Air Force Base)
1. Dr. Junho Choi - Florida Institute of Technology

ELECTRONICS SYSTEMS DIVISION
(Hanscom Air Force Base)
1. Dr. Wilford Fordon - Michigan Technological University

ENGINEERING & SERVICES CENTER
(Tyndall Air Force Base)
1. Dr. Muluneh Azene - Southern University
2. Dr. Robert Douglas - North Carolina State University
3. Dr. Fernando Fagundo - University of Florida
4. Prof. Ervin Hindin - Washington State University
5. Dr. Richard Rice - Clemson University

FLIGHT DYNAMICS LABORATORY
(Wright-Patterson Air Force Base)
1. Dr. Eric Johnson - Virginia Polytechnic Institute & State University
2. Dr. Samuel Kelly - University of Notre Dame
3. Dr. Lakhpat Pujara - Wilberforce University
4. Dr. John Russell - University of Oklahoma
5. Dr. David Wilson - University of Florida
6. Dr. Rama Yedavalli - Stevens Institute of Technology

FRANK J. SEILER RESEARCH LABORATORY
(USAF Academy)
1. Dr. Keith Hanson - Lamar University
2. Dr. Thomas Wiggins - Pennsylvania State University

GEOPHYSICS LABORATORY
(Hanscom Air Force Base)
1. Dr. Francesco Bacchialoni - University of Lowell
2. Dr. Pradip Bakshi - Boston College
3. Dr. Thomas Miller - University of Oklahoma
4. Dr. Stanley Spiegel - University of Lowell
5. Dr. Richard Tipping - University of Nebraska

HUMAN RESOURCES LABORATORY/ADVANCED SYSTEMS DIVISION
(Wright-Patterson Air Force Base)
1. Dr. Mason Bailey - Western Carolina University
2. Dr. Stewart Klapp - California State Universilty

HUMAN RESOURCES LABORATORY/FLYING TRAINING DIVISION
(Williams Air Force Base)
1. Dr. Eugene Galluscio - Northwest Missouri State University
2. Dr. Dean Owen - Ohio State University

PARTICIPANT LABORATORY ASSIGNMENTS (Continued)

HUMAN RESOURCES LABORATORY/PERSONAL RESEARCH DIVISION
(Brooks Air Force Base)
1. Dr. Gary Allen - Old Dominion University
2. Dr. Charles Parsons - Georgia Institute of Technology

HUMAN RESOURCES LABORATORY/TECHNICAL TRAINING DIVISION
(Lowry Air Force Base)
1. Dr. Philip Langer - University of Colorado

LEADERSHIP & MANAGEMENT DEVELOPMENT CENTER
(Maxwell Air Force Base)
1. Dr. Hubert Feild - Auburn University

LOGISTICS COMMAND
(Wright-Patterson Air Force Base)
1. Dr. Donald Byrkett - Miami University

LOGISTICS MANAGEMENT CENTER
(Gunter Air Force Base)
1. Dr. Milton Alexander - Auburn Univerity

MATERIALS LABORATORY
(Wright-Patterson Air Force Base)
1. Dr. Gust Bambakidis - Wright State University
2. Dr. Sherman Brown - University of Illinois
3. Dr. Mark Fugelso - Pennsylvania State University
4. Dr. James Kane - Wright State University
5. Dr. Deborah Medeiros - Pennsylvania State University
6. Dr. David Moroi - Kent State University

ROCKET PROPULSION LABORATORY
(Edwards Air Force Base)
1. Dr. Stephen Lin - North Carolina Central University
2. Dr. Trilochan Singh - Wayne State University

ROME AIR DEVELOPMENT CENTER
(Griffiss Air Force Base)
1. Dr. Richard Absher - University of Vermont
2. Dr. Roger Cavallo - State University of New York
3. Dr. Jerome Knopp - University of Missouri/Rolla
4. Dr. Ronald Remke - University of South Florida
5. Dr. Sam Shanmugan - University of Kansas

ROME AIR DEVELOPMENT CENTER/ELECTRONICS TECHNOLOGY
(Hanscom Air Force Base)
1. Dr. Albert Biggs - University of Kansas
2. Dr. William Filippone - University of Arizona

PARTICIPANT LABORATORY ASSIGNMENT (Continued)

SCHOOL OF AEROSPACE MEDICINE
(Brooks Air Force Base)

1. Dr. Silverio Almeida - Virginia Polytechnic Institute and State University
2. Dr. Willie Bragg - University of Cincinnati
3. Dr. John Enderle - North Dakota State University
4. Dr. Mack Felton - Southern University
5. Dr. Dennis Kern - Universilty of Georgia
6. Dr. Michael Lobb - University of Texas
7. Dr. Troy Nagle - Auburn University
8. Dr. Robert Schlegel - University of Oklahoma
9. Dr. Edward Ward - Alabama A&M University

WEAPONS LABORATORY
(Kirtland Air Force Base)

1. Dr. Francis Jankowski - Wright State University
2. Dr. Letitia Korbly - University of Alabama
3. Dr. Surgounda Patil - Tennessee Technological University
4. Dr. Alexander Stone - University of New Mexico
5. Dr. Edward Walters - University of New Mexico
6. Dr. David Williams - University of Texas

RESEARCH REPORTS
1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Volume I Report Number	Title	Research Associate
1	Sensitivity Based Segmentation And Identification In Automatic Speech Recognition	Prof. Richard G. Absher
2	A Methodology For The Determination Of Input Date Accuracy In The Maintenance Data Collection System	Dr. Milton J. Alexander
3	Assessment of Visuospatial Abilities Using Complex Cognitive Tasks	Dr. Gary L. Allen
4	The Effect Of Modulation Transfer Functions On Flashblindness	Dr. Silverio P. Almeida
5	A Theoretical Evaluation Of The Airfield Pavement Analysis (AFPAV) Finite Element Model	Dr. M. Azene
6	Orbiting Geophysics Laboratory Experiment	Dr. Francesco L. Bacchialoni
7	Trials And Tribulations At The Helmet Mounted Oculomotor Facility	Dr. A. Terry Bahill
8	Feasibility Of Computer Graphics As An Aid To Aircraft Battle Damage Assessment	Dr. M. Gene Bailey
9	Effects Of Magnetic Shear On Lower Hybrid Waves In The Suprauroral Region	Dr. Pradip M. Bakshi
10	A Simple Model For Impurity Photo-Absorption In Silicon	Dr. Gust Bambakidis
11	Feasibility And Implementation Of A Near Field Antenna Range	Prof. Albert W. Biggs
12	Computer Simulation Of Channelized Receivers	Dr. Jack J. Bourquin
13	From Petite Avistrix To USAF Aircrew: Historical Review And Current Status Of Female Fliers	Dr. Willie A. Bragg

Report Number	Title	Research Associate
14	Navier-Stokes Solvers For Ramjet Applications	Dr. Eugene F. Brown
15	Development Of Measurment Technique For Fiber-Matrix Adherence In Brittle-Brittle Composites	Dr. S. D. Brown
16	Force Tracking Proficiency In Operating Aircraft Controls	Dr. L. W. Buckalew
17	Spares Support Using Cannibalization	Dr. Donald L. Byrkett
18	Electrically Compensated Constant Speed Drive	Dr. Jimmie J. Cathey
19	An Expert System For Systems Research In Command And Control	Dr. Roger E. Cavallo
20	Adaptive Kalman Tracking Filter for ARIS System	Dr. Junho Choi
21	Incentive Contracting In Multi-Year Procurement	Dr. David Cleeton
22	Binary Classification And The Subtractive Approach	Dr. Gregory M. Corso
23	Evaluation Of The Non-Destructive Testing Of Airfield Pavements: Wave Propagation Aspects	Dr. Robert A. Douglas
24	AMIS - Acquisition Management Information System Problems And Promises	Dr. Hamed Kamal Eldin
25	Modeling And Tracking Saccadic Eye Movements	Dr. John D. Enderle
26	Dynamic Response Of Doubly Curved Cylindrical Shelter	Dr. Fernando E. Fagundo, J
27	Using Hard Criteria To Evaluate Leadership And Management Development Center Consultations	Dr. Hubert S. Feild

Report Number	Title	Research Associate
28	An Evaluation Of Indices Of Coronary Heart Disease (CHD) In A Diseased Free Population	Dr. Mack Felton, Jr.
29	Electron Transport Calculations Using The Method Of Streaming Rays	Dr. W. L. Filippone
30	Voltammetric Studies Of The Lithium/Vanadium Oxide Electrochemical Cell	Dr. Dennis R. Flentge
31	Integral Principles As Applied To Electromagnetic Propagation	Dr. Wilfred A. Fordon
32	Scaling Laws For Particle Breakup In Nozzle Generated Shocks	Dr. L. J. Forney
33	Oxygen-18 Implantation In GaAs	Dr. Albert J. Frasca
34	Modeling Of Active Neuromusculature Response To Mechanical Stress	Dr. Andris Freivalds
35	The Manufacturing Control Language For Robotic Work Cell	Dr. Mark A. Fugelso
36	Parafoveal Visual Information Processing As A Secondary Task Load	Prof. Eugene H. Galluscio
37	Three-Dimensional Visual Information Processing With Binocular Helmet-Mounted Displays	Prof. Eugene H. Galluscio
38	Shallow Donor And Exciton Binding Energies in Quantum Wells	Dr. Ronald L. Greene
39	Nitro Organic Compounds: A Synthetic Study	Dr. Keith Hansen
40	The Effect Of Certain Image Data Reduction Techniques On Edge Quality	Dr. Donald J. Healy
41	Evaluation Of Gas Chromatographic Methods For Determining Trace Levels Of Trichloroethylene In Water	Prof. Ervin Hindin

Report Number	Title	Research Associate
42	Basic Research Issues In Electromagnetic Rail Launchers With Plasma Driven Projectiles	Dr. Manuel A. Huerta
43	Operational Safety Review Methodology	Dr. Francis J. Jankowski
44	Modeling Localized Stress Fields In Composite Laminates	Dr. Eric R. Johnson
45	Synthesis Of Acetylene Terminated Sulfone (ATS) Candidates	Dr. James J. Kane
46	Mathematical Formulation Of The Shear Layer Over An Open Cavity	Dr. Samuel G. Kelly, III
47	Automating The Numerical Stereo Camera	Dr. Dennis M. Kern
48	Memory And Processing Limits In Decision Making	Dr. Stuart T. Klapp
49	Analysis And Modeling Of A Real-Time Holography System	Dr. Jerome Knopp
Volume II		
50	Numerical Solution Of The Three-Dimensional, Unsteady Euler Equation	Dr. Keith Koenig
51	Displaying Results From Numerical Calculations Using Computer Graphics	Dr. Letitia Korbly
52	Modification Of Current Feedback Strategies: A Text Synthesis Approach	Dr. Philip Langer
53	The Detection Of Hazardous Materials From Spills	Dr. Stephen F. Lin
54	Tentative Identification Awareness Stages Leading To Failure Of Vigilance	Dr. Michael L. Lobb
55	Order Release In An MRP Environment	Dr. D. J. Medeiros

Report Number	Title	Research Associate
56	The Measurement Of Ion-Molecule Reaction Rate Coefficients	Dr. Thomas M. Miller
57	A GaAs FET Computer Model With Applicaions To Magnetic Field Effects	Dr. Michael J. Moloney
58	Flutter Taming By Nonlinear Active Control	Dr. Luigi Morino
59	A Theory Of Photoluminescence For Excitons Bound To Two Types Of Neutral Acceptors In Silicon- A Study Of The Systems Si: (In, Al) and Si: (In, B)	Dr. David S. Moroi
60	Error Analysis Of The Karhunen-Loeve Transform In VCG Signal Processing	Dr. H. Troy Nagle, Jr.
61	Simulation Of Adaptive Networks Based On Rest Principle And Heterostatic Models	Dr. Eugene M. Norris
62	Aircraft Electric Power Generation Using Cascaded Symmetrically Wound Machines	Dr. Thomas H. Ortmeyer
63	Adaptation To Optical Flow Rates During Low Altitude, High Speed Flight	Dr. Dean H. Owen
64	The Robustness Of Unidimensional Item Response Theory Models	Dr. Charles K. Parsons
65	R-Reliable Intervals For The Sum Of Two Independent Normal Or Random Variables	Dr. S. A. Patil
66	Model Reduction Of Control Systems	Dr. L. R. Pujara
67	Measurement Of Levels Of Prostaglandin-E And Prostaglandin-F2α In Various Tissues, Urine, and Plasma Of Rats Given Nonadecafluorodecanoic Acid	Dr. Zahir Qureshi

Report Number	Title	Research Associate
68	A Preliminary Study Of Electromigration In Aluminum And Aluminum-Silicon Films	Dr. R. L. Remke
69	A Study Of Factors Affecting Soot Formation In A Swirl - Stabilized Combustor	Dr. Richard W. Rice
70	Development Of A Mathematical Foundation For Cellular Image Processing	Dr. Gerhard X. Ritter
71	Turbulent Boundary Layers Over Rough Surfaces In Hypersonic Flow	Dr. John M. Russell
72	Conceptualization Of The Dynamic Bahavior Of The Flow-Field In The APL Combustor	Dr. Sarwan S. Sandhu
73	Performance Demand And CNS Depressant Stressor Effects On Spatial Orientation Information Processing	Dr. Robert E. Schlegel
74	Metal Organic Chemical Vapor Deposition-Evaluation Of Current And Possible Future Chemical Systems As Related To The Formation Of Gallium Arsenide	Dr. Eugene P. Schram
75	Software Simulation Requirements For The CVA Program And Other Applications	Dr. K. Sam Shanmugan
76	Afterburning Suppression Kinetics In Rocket Exhaust	Dr. Trilochan Singh
77	On Estimating Probability Of Kill In Weapon Evaluation	Dr. B. D. Sivazlian
78	Methods For Computation Of Certain Diagnostics For The AFGL General Circulation Model	Dr. Stanley L. Spiegel
79	A Differential Geometric Approach To Electromagnetic Lens Design	Dr. Alexander P. Stone
80	The Use Of Experimental Input In Transonic Aerodynamics	Dr. Alfred G. Striz

Report Number	Title	Research Associate
81	A Dynamic Model Of Acceleration Stress Protection IN The Human Aircrew Member	Dr. Patrick J. Sweeney
82	Collision-Induced Absorption In The Far Infrared Spectrum of N_2.	Dr. Richard H. Tipping
83	Photoionization Of Iodine Molecules And Clusters In A Supersonic Molecular Beam	Dr. Edward A. Walters
84	Data Communications Between CDC Cyber 750 and HP 1000	Dr. Kai Wang
85	Isolation Of Pyrogenic Exotoxin C From Staphylococcus Aureus Strains Associated Wtih Toxic-Shock Syndrome	Dr. Edward R. Ward, Jr.
86	Effect Of NDFDA On Fatty Acid Synthesis In Rat Liver	Dr. Thomas E. Webb
87	Hook Interferometry Using A Pulsed Dye Laser	Prof. David Weimer
88	Laser Damage In Films And Plastics	Dr. T. A. Wiggins
89	An Algorithm And Associated Data Structures For Hidden Surface Elimination	Dr. David H. Williams
90	Effort To Produce Metaboric Acid Vapor And Water Vapor Broadening Of Carbon Monoxide Absorption Lines	Dr. Robert E. Willis
91	3-Dimensional Grid Generation With Applications To High Performance Aircraft	Dr. David C. Wilson
92	Time Domain Analysis And Synthesis Of Robust Controllers For Large Scale LQG Regulators	Dr. Rama K. Yedavalli

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

NUMERICAL SOLUTION OF THE THREE-DIMENSIONAL, UNSTEADY EULER EQUATIONS

Prepared by:	Keith Koenig
Academic Rank:	Assistant Professor
Department and University:	Department of Aerospace Engineering Mississippi State University
Research Location:	Air Force Armament Laboratory, Eglin AFB, Munitions Division, Aircraft Compatibility Branch, Aerodynamics Section
USAF Research Colleague:	Mr. Charles J. Cottrell
Date:	August 10, 1982
Contract No:	F49620-82-C-0035

NUMERICAL SOLUTION OF THE THREE-DIMENSIONAL, UNSTEADY EULER EQUATION

by

Keith Koenig

ABSTRACT

The three-dimensional, unsteady Euler equations provide a reasonably general description of inviscid, rotational flow, but are simpler to solve than the Navier-Stokes equations. A computer program, originally developed at AEDC, which solves the general Euler equations for flow past three-dimensional bodies has been implemented on the Eglin AFB, CYBER 176 computer. A document has been produced which discusses in detail this computer program, the fluid mechanics of the problem, and the numerical techniques involved in the solution to the equations. This document serves two purposes: to introduce the principles of computational fluid dynamics to engineers and scientists unfamiliar with this subject, and to provide detailed instructions on the use of this particular computer solution of the Euler equations.

ACKNOWLEDGEMENTS

I would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research, and the Southeastern Center for Electrical Engineering Education for the opportunity they provided me to work at the Air Force Armament Laboratory, Eglin AFB, Florida. I would also like to thank the people of the Aircraft Compatibility Branch for their cooperation and warm hospitality.

A few people deserve special mention. Mr. Charlie Cottrell and Mr. Steve Korn have been most helpful and supportive of my efforts and without them, successful completion of this project could not have been accomplished. Cadet First Class Chris Bogdan provided much needed assistance in working with the computer. Ms. Cynthia Robertson has patiently and quite skillfully typed this long and complex report. I extend my sincere gratitude to each of these people.

I. INTRODUCTION

The equations of motion which describe the general cases of fluid flow are often referred to as the Navier-Stokes equations. If the fluid is assumed to be inviscid then the equations of motion are often called the Euler equations. If the fluid is assumed to be inviscid and the flow is assumed to be irrotational then the flow is termed potential. The full Navier-Stokes equations are a coupled system of second order, non-linear partial differential equations; the Euler equations are a coupled system of first order non-linear partial differential equations and the governing equation for potential flow is a second order, non-linear ordinary differential equation. Describing a flow with the Euler equations is a compromise between the difficulty of solving the equations and the validity or usefulness of the equations. The Navier-Stokes equations can treat virtually any flow problem but are the most difficult to solve. The potential flow equation, although the easiest to solve, is the most limited in its applications. The Euler equations are intermediate in both respects and provide a practical means of solving fairly difficult problems in an economical way. If handled properly the Euler equations may be used to describe three-dimensional, unsteady flow at any Mach number and without assumptions of irrotational (that is, $\nabla \times q = 0$) or isoenergetic (that is, $^{De}/_{Dt} = 0$) conditions. The fact that the Euler equations are not subject to irrotational and isoenergetic assumptions has the important consequence that flows with curved shocks or multiple, intersecting shocks can be treated, which makes the equations applicable to transonic and supersonic flow past complex bodies or combinations of bodies.

II. OBJECTIVES

A computer program was developed at the Arnold Engineering and Development Center to provide solutions of the Euler equations with as few limitations on geometry as possible. This program, designated at AEDC as ARO-1 and

described in Reference 1, solves the three-dimensional, unsteady Euler equations in Cartesian coordinates using a finite volume (volume flux) approach. The basic numerical algorithm is the explicit predictor-corrector scheme of MacCormack (Ref 2). A version of this code has been set up on the CYBER 176 computer at Eglin AFB. This report describes the code (which we call EULER) and its implementation on the Eglin computer so that users unfamiliar with computational fluid dynamics (CFD) programs might be able to apply the code to their own problems. It is strongly recommended that Ref 1 be studied along with this report for a more complete understanding of the program.

III. BASIC CONCEPTS

The Mesh

The Euler equations are differential equations. Numerical solutions to differential equations require approximating continuous functions (such as pressure) and continuous operations (mainly differentiation and integration) by discrete values, differences, and sums done for discrete points in space and time. The discrete points in space form the basis of what is called a computational grid or mesh. This mesh is an essential part of any CFD code. It must be fine enough to resolve important flow details yet coarse enough to keep storage and computational time requirements reasonable. In addition, it should be laid out and numbered in a consistent manner so that the computations can progress smoothly and unambiguously through the mesh. The mesh is also one of the portions of this program that users will need to modify or create on their own for particular problems and so a thorough understanding of this mesh scheme is important.

Basically a mesh or grid should concentrate points near the body and spread them out farther away. Figure 1 shows an example of a very coarse grid about an axisymmetric body such as used in this program. Part a shows the

side view, b shows the front view midway back on the body, and c shows an oblique view of just a portion of the grid. Nomenclature used in the program is also given. The mesh shown here would be referred to as a 10 X 3 X 7 mesh; that is, there are 10 planes perpendicular to the X axis, 3 surfaces enclosing the X axis, and 7 planes radially out from the X axis. The numbering scheme increases the indices with increasing values of X, radial location, and azimuthal location (clockwise from vertical). The subscripts (JX,JY,JZ) are not synonymous with the coordinates (X,Y,Z). For example, in this grid two points with the same JZ may have different Z values.

The network of lines forms six-sided volumes (some, on the X axis degenerate to only five sides) with eight corners or vertices. To refer to a particular volume in the mesh the subscripts for the corner in the most downstream, radially outward, clockwise location are used. The subscripts for this corner are also the largest values of JX, JY, and JZ which any corner in the volume has. For example, the shaded volume in Fig 1c would be designated as volume (5, 3, 2). The physical dimensions of the volumes and their sides are referred to as metrics. Determining these quantities is not as straightforward as might be expected; this will be discussed in more detail later.

In EULER the mesh points are calculated (i.e., grid generated) in one subroutine (currently called CONE) while the metrics are determined in another subroutine, VOLGRD. The subroutine CONE is presently limited to high fineness ratio axisymmetric bodies; users may want or need to write their own grid generation scheme. The subroutine VOLGRD, on the other hand, is quite general and will work for any grid provided that the numbering scheme described above is used.

The Equations

The Euler equations describe flow of an inviscid fluid. For negligible body forces (mainly gravity) these equations can be written (in what is termed conservative form) as

$$\text{Continuity} \qquad \frac{\partial \rho}{\partial t} + \nabla\cdot(\rho q) = 0 \tag{1}$$

$$\text{Momentum} \qquad \frac{\partial(\rho u)}{\partial t} + \nabla\cdot(\rho u q) = -\frac{\partial p}{\partial x} \tag{2}$$

$$\frac{\partial(\rho v)}{\partial t} + \nabla\cdot(\rho v q) = -\frac{\partial p}{\partial y}$$

$$\frac{\partial(\rho w)}{\partial t} + \nabla\cdot(\rho w q) = -\frac{\partial p}{\partial z}$$

$$\text{Energy} \qquad \frac{\partial}{\partial t}\ \rho\left(\tilde{e} + \frac{q^2}{2}\right) + \nabla\cdot\ \rho\left(\tilde{e} + \frac{q^2}{2}\right)q = -\nabla\cdot(pq) \tag{3}$$

Where $q = u\hat{i} + v\hat{j} + w\hat{k}$. In addition an equation of state is required. Here a perfect gas is assumed, so that

$$p = \rho RT \tag{4}$$

$\tilde{e} = c_v T$ = internal energy

$\gamma = c_p/c_v$ = constant

$c^2 = \gamma RT = \gamma p/\rho$ = speed of sound (squared)

With some algebra these equations can be rearranged to all have the same form

$$\frac{\partial G}{\partial t} + \nabla\cdot F = 0 \tag{5}$$

Here G is a vector formed of the dependent flow variables ρ, ρu, ρv, ρw, and e, and F is a matrix of flux quantities such as ρuv (the flux of x momentum in the y direction). More exactly

$$G = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho w \\ e \end{bmatrix} \qquad F = \begin{bmatrix} \rho u & \rho v & \rho w \\ \rho u^2+p & \rho uv & \rho uw \\ \rho uv & \rho v^2+p & \rho vw \\ \rho uw & \rho vw & \rho w^2+p \\ (e+p)u & (e+p)v & (e+p)w \end{bmatrix} \tag{6}$$

The energy e here is actually the sum of the internal and kinetic energy per unit volume

$$e = \rho(\tilde{e} + \frac{q^2}{2}) \qquad q^2 = u^2 + v^2 + w^2 \tag{7}$$

and is related to pressure (proved using equations (4)) by

$$e = \frac{p}{\gamma - 1} + \tfrac{1}{2}\rho q^2 \tag{8}$$

If equation (5) is applied to one of the small stationary volumes, say V, of the mesh it may be written as (see section IV).

$$\frac{\partial G}{\partial t} = -\frac{1}{V}\int_S n \cdot F dS \tag{9}$$

Where n is the outward unit normal to the surface S, enclosing V, and G is the average value of G in the volume. Here this average value is approximated by the value at the center.

Equation (9) is the (schematic) governing equation of Euler flow. It describes three-dimensional unsteady flows, but here is currently being used to calculate steady flows. To do so a set of initial conditions for G (and therefore F) are assumed. These initial conditions are arbitrary but assuming freestream conditions everywhere initially works well. Through the use of equation (9) the time rate of change of G is determined and values of G at later times are found. Time is allowed to march on until the values of G at one time are sufficiently close to the values at the preceding time to consider the process as converged. These converged values of G are then the solution to the problem. The solution is, of course, constrained by boundary conditions. These are, most commonly, uniform freestream conditions far upstream and also radially far away from the body, a specified normal component of velocity at the body surface (usually zero) and some downstream

conditions. These boundary conditions usually insure the uniqueness of the solution.

The program EULER solves equation (9) using MacCormack's explicit predictor-corrector scheme. Details of this scheme as applied to equation (9) are described in the next section. Here it is sufficient to simply observe a few general characteristics of the technique and this equation. Equation (9) provides a means of finding G at a later time knowing F at the present time. The explicit nature of the technique means that at a given instant of time all values of F are known and $\frac{\partial G}{\partial t}$ thus can be found by one equation. (Implicit techniques require solving simultaneous equations.) Because calculations are actually done using finite time steps and values of F known only at certain points (rather than continuous distributions) the time derivative and surface integrals in (9) can only be approximated. The predictor-corrector nature of the technique means that equation (9) is used twice, first to predict values of G and F at a later time, then to take these predicted values and obtain improved or corrected values of G and F at this later time. These corrected values can be shown to be roughly second order accurate in the time and space intervals which is sufficient for many applications.

The Euler equations are for an inviscid fluid and consequently cannot compute boundary layers, for example. In order to increase the usefulness of this code, a boundary layer calculation scheme was incorporated at AEDC into the program (Ref 3). The program can be run, however, with the boundary layer calculations "turned off" and the remaining discussions will be for the inviscid calculations only.

IV. IMPLEMENTATION OF MacCORMACK'S EXPLICIT PREDICTOR-CORRECTOR TECHNIQUE

The equations of motion are schematically written

$$\frac{\partial G}{\partial t} + \nabla \cdot F = 0 \qquad (10)$$

where G represents the flow variables and F represents the flux vectors. We would like to know G at all times throughout space. This would then be the solution to the problem. To find G we proceed as follows.

Suppose at some time t we know G everywhere (either because we assumed some values or have calculated them up to that time). The values of G at a later instant in time can be found by expanding G in a Taylor series about t, that is

$$G(t + \Delta t) = G(t) + \left(\frac{\partial G}{\partial t}\right)_t \Delta t + \tfrac{1}{2}\left(\frac{\partial^2 G}{\partial t^2}\right)_t \Delta t^2 + \ldots \qquad (11)$$

where the subscripts t on the derivatives mean evaluate the derivatives at time t. The larger the time step Δt is, the more terms are needed in (11) to give a good value for $G(t+\Delta t)$. For finite time steps, equation (11) might be approximated by

$$G(t + \Delta t) = G(t) + \left(\frac{\partial G}{\partial t}\right)_{avg} \Delta t \qquad (12)$$

where $\left(\frac{\partial G}{\partial t}\right)_{avg}$ is some effective, or average, value of the time derivative over the interval t to $t+\Delta t$. Finding the new value of G now becomes a problem of finding $\left(\frac{\partial G}{\partial t}\right)_{avg}$. This is done through the use of equation (10). The flux vectors F in (10) are simply algebraic combinations of the flow variables G. If G is known everywhere at time t, then F is also known everywhere at the same instant and equation (10) gives a way to find the first time derivative of G.

To use equation (10) on a grid of discrete points some manipulations are necessary. Rearrange (10) and integrate it over one of the small, stationary grid volumes to give

$$\int_V \frac{\partial G}{\partial t}\, dV = -\int_V \nabla \cdot F dV \qquad (13)$$

Consider the left hand side of (13). Since the volumes are assumed to be stationary, the time derivative can be taken outside. The mean value theorem of calculus can then be applied to the integral. This sequence yields

$$\int_V \frac{\partial G}{\partial t} dV = \frac{\partial}{\partial t} \int_V G dV = \frac{\partial}{\partial t} \left(\tilde{G} \int_V dV\right) = \frac{\partial}{\partial t}(\tilde{G}V) = V\frac{\partial \tilde{G}}{\partial t} \tag{14}$$

where $\tilde{G}$ is the mean value of G in the volume V. If the volume is small then $\tilde{G}$ is approximately the same as the value of G at the volume center and so (14) gives

$$\int_V \frac{\partial G}{\partial t} dV = V \frac{\partial G}{\partial t} \tag{15}$$

Note that equations (13) and (14) are exact mathematically given equation (10) and that equation (11) is also exact (with the proper conditions of G being differentiable and integrable). Equations (12) and (15) introduce approximations. Now consider the right hand side of (13). Use the divergence theorem to change the volume integral to a surface integral. That is

$$\int_V \nabla \cdot F dV = \int_S n \cdot F dS \tag{16}$$

where n is the unit outward normal of S. Use (15) and (16) in (13) and divide through by V to give

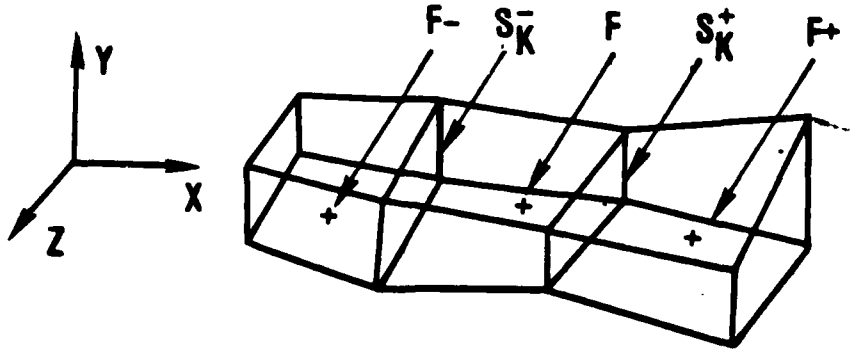
$$\frac{\partial G}{\partial t} = -\frac{1}{V} \int_S n \cdot F dS \tag{17}$$

Equation (17) is now the means by which the first time derivative of G, evaluated at the volume center and at time t, can be found approximately.

Before we lose the forest for the trees let us step back and look at what we are trying to do. Our goal is to find G (t+Δt) knowing G(t). We decided to do this via equation (12). But this equation requires finding $\left(\frac{\partial G}{\partial t}\right)_{avg}$ (which has not even been defined yet). A specific time derivative can be found using equation (17) and so we have a tool to use. However, there is a

catch to this tool. Equation (17) requires an integration of F over the surfaces of a grid volume. But we only know F at discrete points in space, in fact only at the centers of volumes. So we have two problems to solve in order to find $G(t+\Delta t)$, that is, how do we find $(\frac{\partial G}{\partial t})_{avg}$ and how do we do the integration in (17)? These two problems are addressed simultaneously by MacCormack's explicit predictor-corrector technique. Justifying that the steps in MacCormack's method are valid and accurate to a particular order is not that simple but can be done. Let us accept that the technique does work and is essentially second order accurate and worry only with how it is used and that it, at least, is reasonable.

A digression is necessary at this point to review the computational grid. The grid points form the corners of six-sided volumes, three of which are sketched below.



The crosses in the sketch are the volume centers where the flow variables and flux vectors are known, here denoted by F^-, F, and F^+. The two shaded faces S_k^- and S_k^+ are opposing faces of the middle volume to serve as examples.

The integration in equation (17) determines the net flux through the volume in question. Since each volume has three pairs of opposing sides, the integral can be treated as the sum of the integrals over each of the six surfaces, that is

$$\int_S n \cdot F dS = \sum_{k=1}^{3} \int_{S_k^+} n \cdot F dS + \sum_{k=1}^{3} \int_{S_k^-} n \cdot F dS \qquad (18)$$

which is exact. However, F is not known over each surface but only at the volume center points so the individual integrals must be approximated. This approximation can be done in a variety of ways. For example, the flux across S_k^- might be determined using an average of F^- and F, or an average of F^-, F plus other neighboring centers (not shown) or using F^- or F alone. MacCormack's method essentially uses an average of F^- and F but does so indirectly in two steps. The way in which this is done will be seen shortly.

The term predictor-corrector means a solution for $G(t+\Delta t)$ is first predicted, then this prediction is corrected. The quantity $(\frac{\partial G}{\partial t})_{avg}$, defined by

$$(\frac{\partial G}{\partial t})_{avg} = \tfrac{1}{2}((\frac{\partial G}{\partial t})_t + (\frac{\partial \bar{G}}{\partial t})) \tag{19}$$

incorporates this idea; $(\frac{\partial G}{\partial t})_t$ predicts a value and $(\frac{\partial \bar{G}}{\partial t})$ corrects this result. The predicted value of $G(t+\Delta t)$ is given by

$$\bar{G}(t + \Delta t) = G(t) + (\frac{\partial G}{\partial t})_t \Delta t \tag{20}$$

Equations (17) and (18) are used to obtain $(\frac{\partial G}{\partial t})_t$

$$(\frac{\partial G}{\partial t})_t = -\frac{1}{V}(\sum_{k=1}^{3} F \cdot S_k^+ + \sum_{k=1}^{3} F^- \cdot S_k^-) \tag{21}$$

Note that here the flux integrals are approximated by using the flux vectors "behind" the appropriate face. The quantities $\bar{G}(t+\Delta t)$ from (20) can now be used to find predicted values for the flux vectors, that is $\bar{F}(t+\Delta t)$. These new flux vectors are used to find the predicted derivative, which serves as a corrector to $\bar{G}(t+\Delta t)$,

$$(\frac{\partial \bar{G}}{\partial t}) = -\frac{1}{V}(\sum_{k=1}^{3} \bar{F}^+ \cdot S_k^+ + \sum_{k=1}^{3} \bar{F} \cdot S_k^-) \tag{22}$$

Here the flux integrals are approximated by using the flux vectors "ahead of" the appropriate face. Equation (22) is the corrector step and completes the flux averaging problem mentioned earlier.

In principle all the ingredients to find $G(t+\Delta t)$ are now here. First using the known conditions at time t find $(\frac{\partial G}{\partial t})_t$ from equation (21). Substitute this result into (20) to find $\overline{G}(t+\Delta t)$. Use $\overline{G}(t+\Delta t)$ to find $\overline{F}(t+\Delta t)$. Substitute $\overline{F}(t+\Delta t)$ into (22), (22) into (19), and (19) into (12) to yield $G(t+\Delta t)$. In the notation of Ref 1, $G(t) = G^n$, $G(t+\Delta t) = G^{n+1}$, and $\overline{F}(t+\Delta t) = \overline{F}$. Except for a factor ω, equations (20) and (21) here form the first of equations (7) in Ref 1 while equations (12), (19), (20), and (22) here combine to form the second of equations (7) in Ref 1.

V. DISCUSSION AND RECOMMENDATIONS

Time-dependent, predictor-corrector solution techniques of the three-dimensional, unsteady Euler equations can be effectively applied to a rather wide range of practical situations, as demonstrated in References 1 and 3. The program implemented on the Eglin CYBER 176 provides the same capabilities, although somewhat limited by storage constraints. The preceeding sections have presented introductions to three essential aspects of this, or any, CFD program: the equations being solved, the computational grid on which the solution is determined, and the numerical algorithm used to solve the equations. Satisfactory application of this program however, requires users, especially those new to CFD, to thoroughly understand the details of the code. The Appendix to this report contains these details -- the organization and sequence of events, definitions of the variables, and detailed descriptions of the major subroutines. Although some of this information is not absolutely necessary in order to use the program, careful study of the Appendix will allow users to more efficiently apply the code to their particular problem.

The full capabilities of this program can only be determined and exploited if the code is tested and used for a variety of problems; these activities are highly recommended. If the storage constraints are too severe, thought might be given to overlaying or similar space saving techniques. Finally, it is

recommended that a boundary layer calculation and a plotting package be made fully operational, but optional, features of this program to make it a more useful tool.

REFERENCES

[1]Jacobs, J. L. and Kniele, K. R., "Computation of Three-Dimensional, Time-Dependent Flow Using the Euler Equations," Arnold Air Force Station, Tenn., AEDC-TR-80-49, July 1981.

[2]MacCormack, R. W., "The Effect of Viscosity in Hypervelocity Impact Cratering," AIAA Paper No. 69-354, 1969.

[3]Whitfield, D. L, Swafford, T. W., and Jacocks, J. L., "Calculation of Turbulent Boundary Layers with Separation and Viscous-Inviscid Interaction," AIAA Journal, Vol. 19, No. 10, Oct. 1981.

[4]Etkin, B., Dynamics of Atmospheric Flight, John Wiley & Sons, Inc., New York, 1972, pp. 112-117.

[5]Spiegel, M. R., Mathematical Handbook of Formulas and Tables, Schaum's Outline Series in Mathematics, McGraw-Hill, New York, 1968, p. 107.

[6]Anderson, J. D., Modern Compressible Flow, McGraw-Hill, New York, 1982, pp. 312-313.

APPENDIX

A.1 PROGRAM ORGANIZATION

Introduction

The program in its present form on the Eglin computer consists of a main calling program, 20 subroutines and 2 functions. Three of the subroutines and the two functions are used only in the boundary layer calculations and therefore are not of concern now. The invisicid calculations involve considerable looping among the main program and the other 17 subroutines. Extensive use is made of labelled common regions to transfer data among these subroutines. Input to the program is quite minimal. Only four data cards are read although this is a little misleading since the body coordinates are presently given through DATA statements in one of the subroutines and other information is provided through the use of PARAMETER statements. One subroutine is solely for writing output although the use of this subroutine is not necessary; users may produce output as they wish.

Subroutines

A list of the calling program and subroutines follows with very brief descriptions of the main activities of each unit. The list is in the order in which the routines appear in the program and not in the order in which they are called. Some of these subroutines will be described in greater detail in later sections.

HOLDEBR. This is the main calling program. It calls (in chronological order) the subroutine which (1) allows for restart, (2) establishes the initial conditions, (3) generates the grid, (4) computes the grid metrics, (5) initializes certain arrays, (6) prints the grid mesh point locations, (7)

prints the initial conditions, (8) controls the computational cycles, and (9) which stops the program when everything is finished.

START. This subroutine allows for a restart of the program if it is stopped sometime during a calculation. This subroutine also initializes the flux accumulators and establishes the initial boundary conditions and time step.

PVAR. This subroutine computes and prints out surface pressure coefficients. It also controls printing of the grid and flow variables.

PARRAY. This subroutine writes any array passed to it.

CONE. This subroutine contains the body coordinates in DATA statements and computes the mesh point locations. It is restricted to high fineness ratio axisymmetric bodies.

ICEES. This subroutine computes the initial conditions throughout the flowfield.

VOLGRD. This subroutine computes the grid metrics.

BC. This subroutine enforces the boundary conditions (eqn 9, ref 1).

FLUXX, FLUXY, FLUXZ. These subroutines compute the fluxes of mass, momentum and energy through the faces of the grid volume sides in the JX, JY and JZ directions, respectively. Flux accumulators which sum the net flow for the predictor-corrector calculations are also computed.

UPDATE. This subroutine computes the new or updated values of the flow variables (G) using the flux accumulators. This is where equations (7) of Ref 1 explicitly appear.

DTCAL. This subroutine computes the new time step to be used (eqn 8, ref 1).

STEP. This subroutine controls the calls to the FLUX, UPDATE, SMOOTH, and DTCAL subroutines.

BSTEP. This subroutine controls the calls to the STEP and PVAR subroutines.

ERRORS. This subroutine provides a way to stop the program in a controlled way if certain errors occur.

BLAYER. This subroutine computes the boundary layer.

TURB. This subroutine computes boundary layer turbulence.

SIMP. This subroutine computes integrals using Simpson's Rule.

UNINT. This subroutine computes linear interpolations.

SMOOTH. This subroutine computes smoothed or averaged values of the flow variables inside the grid.

FH, DHTS1. These functions are used in the boundary layer calculations.

Inputs

Information is transferred into the program in three ways, through READ, DATA, and PARAMETER statements. For normal operation only four data cards are read. However, if the program is to be run with a restart capability then additional information may be read from a specified device. The restart reads will be indicated but only the normal operation will be considered.

The data cards will be discussed first in the order they are encountered with definitions and typical values of the variables being read.

```
      READ (5,100) KIN, KOUT, TFLAG
100   FORMAT (2I3, E24.0)
```

This read is in SUBROUTINE START.

KIN = number of file to be read if restart is used = 00

KOUT = number of file for output if restart is used = 00

TFLAG = maximum time value permitted if internal clock of computer is used to time the computation = 7.00

```
      READ (5,100) FACT, FSMACH, ALPHA, BETA, PHI
100   FORMAT (8E10.0)
```

This read is in SUBROUTINE ICEES.

FACT = parameter for computational stability (CFL number) = .8

FSMACH = freestream Mach number = 0.8500

ALPHA = freestream flow angle = blank

BETA = freestream flow angle = -00.00 . . . no incidence

= -10.00 . . . 10 deg angle of attack

PHI = freestream flow angle = blank

```
      READ (5,100) NB, NT, NV, FACTX
100   FORMAT (3I5, F20.0)
```

This read is in PROGRAM HOLDEBR. A very important parameter is the product

NB*NT = total number of time steps to be used = 4000 (for example)

NB = number of printouts of surface pressures = 20

NT = (NB*NT)/NB, must be a multiple of 8 = 200

NV = number of times the boundary layer is called = 00

FACTX = parameter for computational stability = blank

```
      READ (5,100)
100   FORMAT (2I3.E24.0)
```

This read is in SUBROUTINE START. A blank card must be the fourth card. That's a good question.

Besides these reads from TAPE5, there are several READ(KIN) statements for restart which appear in SUBROUTINE START.

A second form of input is the DATA statement. The occurrences of DATA are as follows:

In HOLDEBR.

DATA GAM, GM1, GS2, RGM1 /. . ./, where

GAM = γ, GM1 = γ-1, GS2 = γ/2, RGM1 = $(\gamma-1)^{-1}$.

The other two DATA statements set initial values for certain counters and reference numbers.

In CONE.

DATA XB /. . ./ and DATA YB /. . ./ give the x and y coordinates for a particular body of revolution. Other bodies may be used simply by replacing these coordinates with new ones.

In FLUXX, FLUXY, FLUXZ.

DATA KODE /. . ./ provides a way to cycle around the eight corners of a grid volume. These should <u>not</u> be changed.

No other DATA statements occur.

The last input method is the PARAMETER statement. It sets values of certain quantities which then cannot be changed by any later program statement. It must appear in every subroutine using those quantities. Here only one statement is used but it is in almost all of the subroutines.

PARAMETER (NNX = 80, NNY = 32, NNZ = 2, . . .)

This particular statement is for generating an 80 X 32 X 2 grid. Other grids can be formed by replacing these numbers. This must be done in <u>every</u> PARAMETER statement. One major benefit of using the PARAMETER statement is that variables may be dimensioned (in DIMENSION or COMMON statements for

example) with variable sizes, if these sizes are first specified in PARAMETER statements. This saves a tremendous amount of work in this program when grid sizes are changed.

Variables

There are a large number of variables in this program, many of them triply subscripted with dimensions of about (NNX, NNY, NNZ). This leads to large storage requirements. Most of these variables appear in labelled COMMON statements and these labelled COMMON statements appear in most of the subroutines even though the subroutines may actually require only a few of the variables. The variable names are, in many cases, descriptive of the actual quantities they represent which helps in understanding and using the program. Following is a list of the more important variables, their dimension size, definitions, and the subroutines in which they are actually used. This list does not include some insignificant variables and variables which appear exclusively in SMOOTH, LININT, BLAYER, TURB, SIMP, ERRORS, FH, and DHTS1.

A (NNX1, NNY1, NNZ1)	= dummy name for any of the flow variables	PARRAY
ADA	= A·A, magnitude of area vector squared	BC, DTCAL
ALPHA	= freestream flow angle	ICEES
AX, AY, AZ	= average value of the x, y, and z components of the area vectors of opposing faces	VOLGRD

AX1, AY1, AZ1 (NNX, NNY1, NNZ1)	= components of area vector of face with constant JX	VOLGRD, FLUXX, DTCAL
AX2, AY2, AZ2 (NNX1, NNY, NNZ1)	= components of area vector of face with constant JY	VOLGRD, FLUXY, DTCAL, BC
AX3, AY3, AZ3 (NNX1, NNY1, NNZ)	= components of area vector of face with constant JZ	VOLGRD, FLUXZ, DTCAL, BC
BETA	= freestream flow angle, negative of angle of attack	ICEES
CLSTEP	= coefficient of flux summation (eqn (11), Ref 1)	UPDATE
CP2 (NNX1)	= upper surface pressure coefficient (actually slightly above and clockwise of upper surface)	PVAR
CPNNZ (NNX1)	= lower surface pressure coefficient (actually slightly below and counter-clockwise of lower surface)	PVAR

CSV	= (1/Mach number)	DTCAL
DE, DR, DRU, DRV, DRW (NNX1, NNY1, NNZ1)	= flux accumulators for energy, density, and (x, y, z) momentum	START, SMOOTH, FLUX, UPDATE
DELY (NNY)	= spacing between JY = constant grid lines	CONE
DELZ	= spacing between JZ = constant grid lines	CONE
DT, DTM	= time step (eqn (8), Ref 1)	DTCAL
DTSTEP	= coefficient of flux accumulators (eqn (11), Ref 1)	UPDATE, DTCAL
DTSVOL	= coefficient of flux accumulators	UPDATE
DUM	= denominator of eqn (11), Ref 1	DTCAL
DYY	= intermediate variable to find spacing between JX = constant grid lines	CONE

DZ	= spacing between JZ = constant grid lines	CONE
DX1, DY1, DZ1	= components of one face diagonal	VOLGRD
DX2, DY2, DZ2	= components of other face diagonal	VOLGRD
DXC, DYC, DZC	= components of vector between centers of opposing faces	VOLGRD
E (NNX1, NNY1, NNZ1)	= energy	ICEES, BC FLUX, DTCAL
EINIT	= initial value of energy	ICEES
FACT, FACTX	= Courant-Friedrich-Lewy stability parameter	HOLDEBR, ICEES, DTCAL
FF	= $2*(\rho q \cdot A)/(A \cdot A)$, mirror condition, (eqn (9), Ref 1)	BC
FSMACH	= freestream Mach number	PVAR, ICEES
F1, F2, F3, F4, F5 (NNX1, NNY1)	= flux of mass, (X, Y, Z) momentum and energy	FLUXX, FLUXY, FLUXZ

GAM	= γ, ratio of specific heats	ICEES, PVAR
GM1	= γ-1	ICEES
GS2	= γ/2	DTCAL
INC	= counter, number of times STEP is called	STEP
JJ, JS, JT	= indices for predictor-corrector	FLUX
JX, JY, JZ	= primary indices	all sorts of places
JX1, JX2, JY1, JY2 JZ1, JZ2	= indices for predictor-corrector	FLUX
KIN	= input file number for restart	HOLDEBR, START
KODE	= code for predictor-corrector	FLUX
KOUNT	= counter for time steps	PVAR, BC, STEP
KOUT	= output file number for restart	HOLDEBR, START

LIMIT	= counter limit for calls to STEP	STEP
NB	= number of printouts	HOLDER
NEX, NEY, NEZ	= grid size -1	all sorts of places
NNX, NNY, NNZ	= grid size	all sorts of places
NNX1, NNY1, NNZ1	= grid size +1	all sorts of places
NT	= frequency of printouts	HOLDEBR, BSTEP
NV	= number of calls to BLAYER	HOLDEBR
OMEGA	= artificial viscosity	UPDATE
OMTO	= coefficient of flux accumulators	UPDATE
P (NNX1, NNY1, NNZ1)	= pressure	PVAR, BC, FLUX, START UPDATE, DTCAL, SMOOTH
PHI	= freestream flow angle	ICEES
PI	= π	CONE
PINIT	= initial value of pressure	ICEEES

QDA	= q•A, volume flux	BC, FLUX
R (NNX1, NNY1, NNZ1)	= density	START, ICEES, BC, FLUX UPDATE, DTCAL, SMOOTH
RGM1	= $(\gamma-1)^{-1}$	DTCAL
RQDA	= $\rho q \cdot A$	DTCAL
RQDQ	= $\rho q \cdot q$	SMOOTH, UPDATE, DTCAL
RQDRQ	= $\rho q \cdot \rho q$	DTCAL
RU, RV, RW (NNX1, NNY1, NNZ1)	= (X, Y, Z) momentum	START, ICEES, BC, FLUX UPDATE, DTCAL, SMOOTH
RVVIS	= viscous y momentum	BC
RUINIT, RVINIT, RWINIT	= initial (X, Y, Z) momentum	ICEES
S (NNX1)	= distance along surface	PVAR
TFLAG	= internal computer time limit	START, STEP
THETAZ	= angular increment in JZ planes	CONE

TIME	= current flow time	PVAR, PARRAY, UPDATE
VOL (NNX1, NNY1, NNZ1)	= grid volume size	VOLGRD, UPDATE, DTCAL, SMOOTH
VOLMIN	= minimum grid volume size	VOLGRID
X, Y, Z (NNX, NNY, NNZ)	= grid coordinates	PVAR, CONE, VOLGRD
XB(60), YB(60)	= body coordinates	CONE
XC, YC, ZC (NNX)	= coordinates of face centers	VOLGRD
XDUM	= denominator of eqn (11), Ref 1	DTCAL
YYT	= intermediate variable to find spacing between JX = constant grid lines	CONE

A.2 PROGRAM SEQUENCE

The major calls, loops, and computations, in the order they occur are described below. This sequence is for a run without the boundary layer calculations, the internal clock, or the restart option, and is for the analysis of one body started impulsively with freestream conditions everywhere. The activities described are undertaken in the underlined program or subroutine immediately above the description.

The input cards will be taken as

KIN = 0 ⇨ no restart

KOUT = 0

TFLAG = 7.00

FACT = .8

FSMACH = 0.85

ALPHA = 0.0

BETA = 0.0

PHI = 0.0

NB = 20 } ⇨ 4000 time steps with a printout every

NT = 200 } 200 time steps

NV = 0 ⇨ no call to boundary layer

FACTX = 0.0

These are typical values for a run without the boundary layer or restart.

HOLDEBR

calls START

reads KIN, KOUT, TFLAG

if KIN = 0 returns to HOLDEBR

calls ICEES

reads FACT, FSMACH, ALPHA, BETA, PHI

computes initial values of R, RU, RV, RW, E

returns to HOLDEBR

calls CONE

body coordinates XB, YB entered via DATA

computes (X, Y, Z) grid points (calls LININT to help)

returns to HOLDEBR

calls VOLGRD

computes AX1, AY1, AZ1, AX2, AY2, AZ2, AX3, AY3, AZ3, VOL

returns to HOLDEBR

calls ENTRY MIDDLE in START

computes P

initializes flux accumulators DR, DRU, DRV, DRW, DE

calls BC

computes initial values of ADA, QDA, R, E, P, FF, RU, RV, RW at phantom and downstream points

returns to MIDDLE

if KIN = 0 and KOUT = 0 returns to HOLDEBR

calls ENTRY PGEOM in PVAR

writes grid coordinates (X, Y, Z)

returns to HOLDEBR

calls PVAR

calls PARRAY

writes R, RU, RV, RW, EP

returns to PVAR

computes and writes initial values of CP2 and CPNNZ

returns to HOLDEBR

reads NB, NT, NV, FACTX

At this point a nested loop process begins which, for the most general situation, is controlled by the values of NB, NT, NV, and TFLAG. For the sequence described here where the internal computer clock and boundary layer are not being used then only NB and NT control the process.

HOLDEBR calls BSTEP (NB times)

BSTEP calls STEP (NT times)

STEP increments KOUNT by 1

STEP calls FLUX and UPDATE (2 times for each)

FLUX computes DR, DRU, DRV, DRW, DE

UPDATE computes TIME, R, RU, RV, RW, E, P,

DR, DRU, DRV, DRW, DE

UPDATE calls BC

BC computes R, E, P, RU, RV, RW at boundaries

STEP calls SMOOTH after every 16 calls to FLUX and UPDATE

SMOOTH computes DR, DRU, DRV, DRW, DE, R, RU, RV, RW, E, P

SMOOTH calls BC

BC computes R, E, P, RU, RV, RW at boundaries

STEP calls DTCAL after every 128 calls to FLUX and UPDATE

DTCAL computes DTSTEP

BSTEP calls PVAR

PVAR computes and writes CP2 and CPNNZ

When this loop sequence is finished KOUNT will equal the product NB*NT, there will be NB printouts of CP2 and CPNNZ, and the program will be back in HOLDEBR. For the case of no internal clock and no restart the remaining steps end the program.

HOLDEBR calls ENTRY FINISH in START

if KOUT = 0, returns to HOLDEBR

STOP

END

A.3 SUBROUTINES

Cone

This subroutine generates a grid as shown in Fig 1 for slender axisymmetric bodies. In its original form CONE generated an 80X32X2 grid for a body of revolution mounted on a wind tunnel sting support. This body has been kept throughout the program development and still appears in the program via the DATA in CONE. The original grid generation equations also remain in the program but have been "commented out." This allows a user to see what a good grid scheme looks like.

One difficulty in using this program on the Eglin computer is limited core storage. An 80X32X2 grid approaches the maximum core storage which is available; however, such a grid is only good for zero incidence flow. If the body is put at some angle to the flow, then the last number in the grid size (the number of azimuthal or JZ stations, NNZ) must be increased and a corresponding decrease is necessary in the other two grid dimensions (NNX and NNY) to keep within the machine storage limits. It is in determining the grid points in the JX and JY directions that great care must be taken. If the number of JX and JY points are decreased without changes to the grid equations then the resulting grid will be very inadequate. So, it is necessary to adjust these equations to handle a variable number of JX and JY points (that is variable NNX and NNY). Attempts to do this in a general way have not been very successful and this is a weak point in the program at present.

The original grid scheme creates JX lines which are perpendicular to the X axis. On the body, the lines are evenly spaced. Ahead and behind the body the line spacing increases with distance from the body. A similar situation exists for the spacing of the JY lines. These lines are closely spaced near the body and increase in spacing away from the body. The original scheme used in CONE for an 80X32X2 grid is based on the following equations. The increment between JY lines is

DELY(JY) = 0.0025*JY**1.2

Thus, for the first vertical grid line (JX = 1, JZ = 1) we have

$$Y(JY) = 0.0025^* \sum_{j=2}^{JY} j^{1.2} + Y(1) \qquad (1)$$

Where JY > 1, Y(1) = 0.0, Y(NNY = 32) = 2.4078

For JX lines not on the body and NNX = 80,

$$X(JX) = -1.25^*TAN(1.25 - 2^*1.25^*\frac{JX}{80})/TAN(1.25) + 0.4384878 \qquad (2)$$

For JX lines on the body

$$X(JX) = X(JX - 1) + \frac{1.049333}{57} \qquad (3)$$

If we desire a grid of smaller dimensions, that is NNX < 80 and NNY < 32, but still desire to retain the characteristics of the grid generated by equations (1), (2), and (3) above, we might try to mimic these equations but make them functions of NNX and NNY. There are certainly many ways to do this; one technique tried here which works reasonably well follows.

Generalizing equation (1) we write

$$Y(JY) = (COEFF)^*\sum_{j=1}^{JY} j^{(POW)}$$

We require (for no good reason) that Y(NNY) = 2.4078 and $\sum_{j=1}^{NNY} j(POW) = \frac{2.4078}{0.0025}$;

the second condition keeps COEFF = 0.003. If we assume that $12 \leq NNY \leq 32$ and play around with POW we find that

$$POW = 0.001875^*(NNY)^2 - 0.1225^*NNY + 3.2$$

works fairly well. In deriving this we use

$$\sum_{i=1}^{n} i^p = \frac{n^{p+1}}{p+1} + \frac{1}{2}n^p + \frac{1}{12}pn^{p-1} - \frac{1}{120}p(p-1)(p-2)n^{p-3} + \ldots$$

which can be found in standard math tables such as Ref 5. The result of all this gives the coding

```
POW = 0.001875*NNY**2 - 0.1225*NNY + 3.2
SUM = (NNY**(POW + 1.))/(POW + 1.) + 0.5*NNY**POW + POW*NNY**(POW - 1.)/12
      - POW*(POW - 1.)*(POW - 2.)*NNY**(POW - 3.)/720.
COEFF = 2.4078/SUM
DELY(JY) = COEFF*JY**POW
Y(JX,JY,1) = Y(JX,JY-1,1) + DELY(JY)
```

Equation (2) can be generalized simply by replacing the 80 by NNX. Equation (3) is generalized by replacing the 57 by DX where

$$DX = \frac{57}{80} * NNX$$

The generalizations just described have been incorporated in some versions of EULER and also in a test program called XYZ7, which generates and plots the grid. This grid is reasonable, but the JX lines off the body are not far enough away, especially JX = 1 and JX = NNX, when NNX < 80.

Another attempt at a generalized grid scheme presently is in the test program GENGRID. The JY increments are found by

```
DELY(JY) = 0.0025*JY**POW                    where

POW = ALOG(2000.0/FLOAT(NNY1))/ALOG(FLOAT(NNY))
```

The equation for POW is obtained by first requiring that

$0.0025 * \sum_{j=1}^{NNY} j^p = 2.4078$ just as in the original scheme.

So this says $\sum_{j=1}^{NNY} j^p = \frac{2.4078}{0.0025} = 963.$

This represents an equation for p as a function of NNY, but it is non-linear and difficult to solve. An approximate solution can be obtained using equation (6) but keeping only the first two terms (and also rounding 963 to 1000). Therefore

$$\sum_{j=1}^{NNY} j^p \approx \frac{NNY^{p+1}}{p+1} + \tfrac{1}{2} NNY^p = 1000$$

or

$$NNY^p \left(\frac{NNY}{p+1} + \tfrac{1}{2}\right) = 1000.$$

Continuing our approximations take $\frac{1}{p+1} \approx \tfrac{1}{2}$ since p should not be too different from 1.

This gives us $NNY^p \tfrac{1}{2}(NNY + 1) = 1000$

or $$p = POW = \ln\left(\frac{2000}{NNY+1}\right) / \ln(NNY)$$

which is the desired result.

The JX lines are found by a procedure quite different from the original scheme, although the character of the original grid is retained. We wish to have JX lines closely spaced near the body and farther spaced away. We also impose the following conditions:

X(1) = -1

X(0.15 NNX) = 0 (nose of body)

X(0.8625 NNX) = 1 (tail of body)

X (NNX) = 2

If we let X be given by a cubic equation then we have four equations in four unknowns, that is

$$X = C_1 + C_2\ JX + C_3\ JX^2 + C_4\ JX^3$$

Finding the constants C_1, C_2, C_3, C_4 which really are functions of NNX gets a little messy (see the coding in the test program GENGRID) but this scheme works, after a fashion. The cubic equation is used for JX lines off the body and constant spacing is used on the body, just as the original scheme.

The original grid scheme in CONE proceeds as follows. First, the body coordinates are non-dimensionalized by the body length. Then the JX lines are calculated and the spacing between the JY lines (but not the lines themselves) are determined. The body coordinates are not used in either of these calculations, only the body length. With the JX lines determined, that is knowing their X locations, the corresponding Y locations on the body are determined by linearly interpolating between the input body coordinates. SUBROUTINE LININT is used for this operation. The X axis and the surface of the body are now taken as the JY = 1 line and the JY spacings previously calculated are used to find all the other JY lines. Finally, the grid at other azimuthal stations (JZ) is found by stepping around the body in specified angular amounts (usually a constant angle). The result of this is a network of NNX X NNY X NNZ points.

Certain features of the grid should be maintained for any grid dimensions. About 70% of the JX lines should be on the body. The first

(JX=1) and last (JX=NNX) lines should be at least one body length ahead and behind the body respectively, i.e., X < -1 and X > 2. The first JY line off the body (JY=2) should be spaced about 0.0025 body lengths away. The last JY line (JY=NNY) should be about two body lengths away. The number of JZ lines depends on the body shape and flow inclination. For an axisymmetric body at 0 incidence two JZ lines are necessary with a spacing of about 20°. For an axisymmetric body at angle of attack the last JZ line (JZ=NNZ>2) should be at 180° and NNZ should be large enough to resolve flow details. Other body shapes will require individual consideration.

ICEES and Freestream Conditions

Quoting from Ref 1, "All variables are dimensionless, with the reference conditions usually taken to be free-stream density and sound speed, the latter given by

$$c = (\gamma \frac{p}{\rho})^{\frac{1}{2}} \quad " \qquad \text{(eqn 4, ref 1)}$$

The variables are R, P, E, RU, RV, and RW. Since these are dimensionless, then, for example

$$R = \frac{\rho}{\rho_\infty} \text{ and therefore } R_\infty = \frac{\rho_\infty}{\rho_\infty} = 1.$$

Similarly, if C is the non-dimensional sound speed, then

$$C = \frac{c}{c_\infty} \text{ and therefore } C_\infty = \frac{c_\infty}{c_\infty} = 1.$$

Since ρ_∞ and c_∞ are the reference values (by definition) then

$$P = \frac{p}{\rho_\infty c_\infty^2} \quad \text{(and not } \frac{p}{p_\infty} \text{ as one might guess).}$$

But from eqn 4, ref 1, $p = \frac{1}{\gamma}\rho c^2$, so

$$P = \frac{1}{\gamma}\frac{\rho c^2}{\rho_\infty c_\infty^2} = \frac{1}{\gamma}RC^2 \text{ and } P_\infty = \frac{1}{\gamma}R_\infty C_\infty^2 = \frac{1}{\gamma}.$$

From eqn 3, ref 1, we have $e = \frac{p}{\gamma-1} + \frac{1}{2}\rho V^2 = \frac{p}{\gamma-1} + \frac{1}{2}\gamma p M^2$

$$E = \frac{e}{\rho_\infty c_\infty^2} = \frac{1}{\gamma-1}\frac{p}{\rho_\infty c_\infty^2} + \frac{\gamma}{2}\frac{p}{\rho_\infty c_\infty^2} M^2 \text{ and } E_\infty = \left(\frac{1}{\gamma-1}\right)\frac{1}{\gamma} + \frac{1}{2}M_\infty^2 .$$

Summarizing these results, the freestream non-dimensional flow is described by

$$R_\infty = 1,\ C_\infty = 1,\ P_\infty = \frac{1}{\gamma},\ E_\infty = \frac{1}{\gamma}\left(\frac{1}{\gamma-1}\right) + \frac{1}{2}M^2$$

If we let q be the dimensioned velocity vector and q be its magnitude, then

$$q = (u^2 + v^2 + w^2)^{\frac{1}{2}},\ M = \frac{q}{c}$$

Also, $U = \frac{u}{c_\infty}$, $V = \frac{v}{c_\infty}$, and $W = \frac{w}{c_\infty}$ so that

U_∞, V_∞, and W_∞ are simply the (X, Y, Z) components of M_∞. This completes the description of the freestream.

The initial conditions are taken to be freestream conditions everywhere. These are established in ICEES. There are five equations of interest in ICEES, which evaluate initial values of RU, RV, RW, P, and E, that is,

RUINIT = $R_\infty U_\infty$ = x component of M_∞ (since $R_\infty = 1$)

RVINIT = $R_\infty V_\infty$ = y component of M_∞

RWINIT = $R_\infty W_\infty$ = z component of M_∞

EINIT = $E_\infty = \left(\frac{1}{\gamma-1}\right)\frac{1}{\gamma} + \frac{1}{2}M_\infty^2$

PINIT = $P_\infty = \frac{1}{\gamma}$ (the equation in ICEES reduces to this).

The components of M_∞ are found from an Euler angle type approach (see, for example, Ref 4). The orientation is slightly unconventional, however. The

coordinate frame remains fixed to the body and the incoming flow is put at an angle. For the coordinate frame as established in Fig 1, a positive angle of attack (nose up) is given by BETA equal to a negative angle. All angles in ICEES are in degrees. ALPHA and PHI refer to yaw and roll angles; their sense is best determined by plugging numbers into the equations and reasoning out the results.

VOLGRD

This subroutine computes the area vectors of the grid volume element sides and the volume of the elements. These computations of the metrics, as they are often called, are done in what may seem to be a rather strange fashion. Before the actual details of the calculations are described, perhaps a few words on the philosophy of the scheme are in order.

This subroutine was written to be able to handle almost any conceivable grid, as long as the grid is consistently numbered. In many cases (such as in this program) the grid elements are some sort of regular geometric volumes with well defined flat sides. However, this is not always the case. It is quite possible to form a grid such that the corners of a volume side are not all in the same plane and consequently the surface is not flat, as depicted in Figure 2. This introduces an essential difficulty. When the side is flat, its area and normal vector (which together give the area vector) can be found unambiguously and exactly (assuming the corners are connected by straight lines). But, if the four corners of a side are not coplanar then the area is not well defined since only the four corners are actually known. Even if the surface can be analytically described it will have a normal vector which varies in direction over the surface. For a numerical solution it is highly desirable to have each surface of a volume element defined by a single, constant area vector. This means the nonplanar surface must be represented

by an effective area vector which must be determined solely in terms of the corner coordinates. In addition, the method for doing this should yield the exact answer when the surface is flat. Finally, the method should be capable of properly handling surfaces that degenerate to a line or point (as happens in almost any grid scheme). The scheme used in VOLGRD satisfies all these requirements.

Determining the volume of an element has similar difficulties. The volume of a flat sided element can be found unambiguously and exactly (although the calculations may be quite tedious). For nonplanar surfaces the element is poorly defined and the volume becomes difficult to compute. For numerical solutions this computation needs to be simplified and based only on the corner coordinates. Also, the result must be exact for flat sided elements. The technique in VOLGRD satisfies these requirements as well.

A representative volume is shown in Figure 3. With the corners labelled as shown, this volume would be referred to by the subscripts (JX, JY, JZ). The surfaces of the element are surfaces on which the values of JX, JY, or JZ are constant and are denoted as A1, A2, and A3 respectively, there being two of each of these for any element. The two A1 surfaces have been singled out here to serve as examples. The calculations to be described are repeated for the other two surface pairs. All surfaces are denoted by the corner having the largest subscripts; so in this example the two A1 surfaces are referred to by (JX-1, JY, JZ) on the left and (JX, JY, JZ) on the right.

Each surface has two diagonals which can be treated as vector lengths. For surface A1 (JX, JY, JZ) the diagonal vectors are

$$D1 = (X(JX,JY,JZ) - X(JX,JY-1,JZ-1))\hat{i} + (Y(JX,JY,JZ) - Y(JX,JY-1,JZ-1))\hat{j}$$
$$+ (Z(JX,JY,JZ) - Z(JX,JY-1,JZ-1))\hat{k} = DX1\hat{i} + DY1\hat{j} + DZ1\hat{k}$$

$$D2 = (X(JX,JY-1,JZ) - X(JX,JY,JZ-1))\hat{i} + (Y(JX,JY-1,JZ) - Y(JX,JY,JZ-1)\hat{j}$$
$$+ (Z(JX,JY-1,JZ) - Z(JX,JY,JZ-1))\hat{k} = DX2\hat{i} + DY2\hat{j} + DZ2\hat{k}$$

The area vector of surface A1(JX,JY,JZ) is defined to be one-half the cross product of the two diagonal vectors, that is

$$A1(JX,JY,JZ) = \tfrac{1}{2}(D1XD2) = \tfrac{1}{2}((DY1*DZ2-DZ1*DY2)\hat{i} + (DZ1*DX2-DX1*DZ2)\hat{j}$$
$$+ (DX1*DY2-DY1*DX2)\hat{k}) = AX1\hat{i}+AY1\hat{j}+AZ1\hat{k}$$

which is exact for flat surfaces with straight lines connecting the corners. (Try it with a trapezoid, for instance.) This vector is also shown on Figure 3. Similarly the diagonal vectors and area vector for side A1(JX-1,JY,JZ) are found.

Each side also has a center which is defined to be the average of the corner coordinates. For example, the center of A1(JX,JY,JZ) has an X value of

$$XC(JX,JY,JZ) = \tfrac{1}{4}(X(JX,JY,JZ) + X(JX,JY,JZ-1) + X(JX,JY-1,JZ-1)$$
$$+ X(JX,JY-1,JZ))$$

The centers of opposing surfaces, here for example A1(JX-1,JY,JZ) and A1(JX,JY,JZ), can be connected by a line which also can be expressed as a vector. That is

$$DC1 = (XC(JX,JY,JZ) - XC(JX-1,JY,JZ))\hat{i} + (YC(JX,JY,JZ) - JC(JX-1,JY,JZ))\hat{j}$$
$$+ (ZC(JX,JY,JZ) - ZC(JX-1,JY,JZ))\hat{k}$$
$$= DXC\hat{i} + DYC\hat{j} + DZC\hat{k}$$

The two area vectors of opposing faces and the vector between the centers of these faces can be used to compute a volume (not necessarily the true volume of the element). For the example here this calculation is

$$VOL = \frac{1}{2}(A1(JX,JY,JZ) + A1(JX-1,JY,JZ)) \cdot DC1$$

that is, the dot product between the average of the opposing face area vectors and the vector connecting the centers of these faces gives a volume. For a rectangular parallelpiped this calculation would indeed be the true volume. Any pair of opposing faces could be used for a rectangular parallelpiped and the result would be the same. Or the average of this process for the three pairs could be used to recover the true volume of a rectangular parallelpiped. For an arbitrary element, the volume is <u>defined</u> by the average for the three pairs. That is,

$$VOL = \frac{1}{3}((A1(JX) + A1(JX-1))\cdot DC1 + (A2(JY) + A2(JY-1))\cdot DC2 + (A3(JZ) + A3(JZ-1))\cdot DC3)$$

where the subscript notations for the area vectors have been shortened.

These definitions of area vectors and volumes are certainly not exact for arbitrary elements but for small volumes are not too far off. For regular elements they are exact and they are computationally efficient. These schemes are similar to those used in many modern numerical computational programs.

A final comment on the nomenclature used in the preceding discussion is necessary. The equations appearing in VOLGRD are in terms of the components of the various vectors; the vectors themselves never occur. The example, the symbol D1 has been used in this discussion for one of the diagonals. In VOLGRD only the components of D1, denoted DX1, DY1, DZ1, are used. Similarly

A1 never appears but AX1, AY1, and AZ1 are used, and so on. Hopefully this difference in notation will not cause any confusion.

BC

This subroutine enforces the boundary conditions on the computed flow. Typically boundary conditions for a body in an unbounded freestream are freestream conditions far away from the body and a specified normal velocity at the body surface (usually zero). Implementing conditions like these on a computational mesh of finite extent requires that these conditions be handled in a special way. The technique used here is based on the concept of phantom points, that is points which are physically outside of the computational mesh.

Phantom points are the centers of volume elements which lie just outside of the computational mesh. Refer back, for a moment, to Fig 1, which shows a (coarse) grid for an axisymmetric body which would be useful for zero incidence or angle of attack problems (hence the JZ lines go half way around the body). The first volume elements in the mesh to be encountered as one goes downstream are those having JX=2 since the volume element is designated by a downstream corner (also see Fig 4a). If one referred to a volume (1, JY, JZ) this would be immediately upstream of the mesh. The center of this volume is a phantom point. Now, referring to Figs 4b and 4e, a volume denoted by (JX,JY,1) would be just to the left (or counter clockwise) of the grid volumes above the body since the volume element is designated by a clockwise corner. The center of this volume is also a phantom point. In a similar manner, the centers of volumes (JX,JY,8), where NNZ + 1 = 8, are phantom points just clockwise of the lower grid volumes (Figs 4b and 4d). The centers of volumes (NNX + 1, JY, JZ) are phantom points just downstream of the last streamwise volumes (Fig 1a). Finally the centers of volumes (JX,1,JZ) are phantom points just inside the body, since volumes are designated by radially outward corners (Fig 4c).

The phantom points are used in the following way. At planes of symmetry (for example lines JZ = 1 and JZ = 7 in Fig 1) and body surfaces (JY = 1) mirror conditions are applied. Normally on a solid body surface or on a plane of symmetry we would simply require that $q \cdot n = 0$ (q = velocity vector, n = normal to surface). However in the formulation of this program we do not know conditions on surfaces but only at the centers of volume elements. Consequently, we only know conditions slightly off the body or to one side of a plane of symmetry. Mirror conditions say that the velocity vector at the phantom point inside the body surface or on the other side of the symmetry plane has a normal component equal and opposite to the normal component of velocity at the neighboring true volume center. One might say that the phantom point velocity vectors are 'mirror images' of the neighboring true velocity vectors, hence the name mirror conditions. For small volume elements these conditions, then, essentially impose a zero normal velocity on the surface in question. Implementing the mirror condition leads to eqn (9) of ref 1,

$$q_{\text{phantom point}} = q_{\text{true}} - 2\left(\frac{S \cdot q_{\text{true}}}{S\ S}\right)S$$

where S is the area vector of the surface in question. In SUBROUTINE BC these mirror conditions are imposed at the phantom points (JX, 1, JZ), (JX, JY, 1) and (JX, JY, NNZ+1). The density, energy, and pressure at these same phantom points are determined by imposing a zero normal gradient condition, that is, R, E, and P do not change from the phantom point to the neighboring volume center. Thus we have R(JX,1,JZ) = R(JX,2,JZ), R(JZ,JY,1) = R(JX,JY,2) and R(JX,JY,NNZ+1) = R(JX,JY,NNZ) and so on.

At the downstream boundary (JX = NNX) the zero normal gradient condition is imposed on all the flow quantities R, E, P, and RU, RV, RW. Thus we have R(NNX + 1, JY, JZ) = R(NNX,JY,JZ) and so on. This condition may not be physically true, especially if the downstream boundary is not far away from

the body, but is has been found to work virtually as well as more correct (and much more complicated) outflow conditions and thus has been retained in the program.

The boundary conditions used in BC never directly impose freestream conditions on any of the computational points. The initial conditions are freestream conditions but these are allowed to evolve in time. This is as it should be since the flow boundaries may actually not be far enough away to have freestream conditions. The mirror and zero normal gradient conditions are sufficient for this solution.

UPDATE

This subroutine essentially evaluates equation (7) of Ref 1 (really its alternate form, eqn (11)). This equation, which is in two parts, is the net result of applying the predictor-corrector technique to this problem. In a very schematic way, both parts of equation 7 (or 11) look like

$$G_{new} = G_{old} - \text{constant} * \sum F \cdot S \qquad (1)$$

The term G represents any of the quantities R, RU, RV, RW, E, and the term F represents the fluxes corresponding to these quantities (see Section III of the main paper). The quantity F·S is referred to as the flux accumulator.

Suppose G = R. The flux accumulator for R is the variable DR. The constant in (1) involves Δt and volume, that is DT and VOL, and is called DTSVOL. So, if we write equation (1) above for G = R we obtain

$$R(JX,JY,JZ) = R(JX,JY,JZ) - DTSVOL(JX) * DR(JX,JY,JZ) \qquad (2)$$

which is precisely one of the equations in SUBROUTINE UPDATE. This equation updates the old value of R by the net mass which has entered the volume during the time interval Δt, hence the subroutine name.

Now, equation (1) represents either part of eqn 7, Ref 1; that is, the predictor and the corrector steps have the same form. Equation (2) above is valid for either step also. Which step is actually being computed at some instant in the program depends on a complicated manipulation of indices and counters which takes place in several subroutines. These manipulations are beyond the scope of the present discussion, other than to observe that they occur.

It might be noticed in UPDATE that a function CVMGT appears. This function is a command which is recognized by CRAY computers but not by the CYBER 176. These statements have been commmented out and the equivalent of them written below them. These statements perform one of the many activities required to properly step through the mesh and the predictor-corrector sequence

DTCAL

This subroutine determines the time step, Δt, to be used in the predictor-corrector calculations. It would be desirable for this time step to be as large as possible so that a solution could be obtained quickly and thus economically. However, the computational scheme is unstable for time steps greater than some critical value. By unstable we mean the solution diverges as time increases rather than converges as we desire. The critical value is determined by the Courant-Friedrich-Lewy stability criterion (Ref 6) which says that the maximum allowable time step is the minimum of

$$\Delta t = \frac{V}{|q \cdot S| + c|S|} \qquad \text{throughout the mesh,}$$

where V is the volume of an element, S and q are area and velocity vectors, respectively, and c is the local speed of sound. This criterion essentially says that the time step can be no greater than the shortest time it takes for an acoustic signal to travel from one volume center to the next, which is certainly physically reasonable.

DTCAL proceeds by computing at a given time

$$\frac{V}{|q \cdot S| + c|S|}$$ for every volume in the mesh and

searching for the minimum value. It is interesting and instructive to look at a few details of how this quantity (called DT) is actually calculated.

Careful study of DTCAL shows that, in the notation of this program but leaving off the subscripts, DT is determined by the expression

$$DT = \frac{VOL*R}{|RQDA| + (RQDRQ*ADA*\frac{GS2}{E/P - RGM1})^{\frac{1}{2}}}$$

where RQDA $= \rho q \cdot A$

RQDRQ $= \rho q \cdot \rho q = \rho^2 q^2$

ADA $= A \cdot A = A^2$

and GS2/(E/P - RGM1) $= 1/M^2$ if the definitions of GS2, RGM1, E, and P are used. This expression for DT can then be written as

$$DT = \frac{VOL*\rho}{\rho|q \cdot A| + (\rho^2 q^2 A^2/M^2)^{\frac{1}{2}}}$$

But $M = q/c$ so this reduces to

$$DT = \frac{VOL}{|q \cdot A| + CA}$$

which is the same as the equation for Δt.

FLUX

The subroutines FLUXX, FLUXY, and FLUXZ (collectively referred to as FLUX) compute the fluxes of the various flow variables across each pair of surfaces of a volume element. Thus, FLUXX computes the flux across the two A1 surfaces, FLUXY computes the A2 fluxes, and FLUXZ determines the transport across the two A3 surfaces. These routines also increment the flux accumulators, which symbolically are F•S, with appropriate care for the signs. The three routines are identical in form and differ only in the subscripts and variables used.

The computations in FLUX are straightforward. Central to these calculations is the definition of the flux of something:

$$\text{flux} = \int_S \left(\frac{\text{something}}{\text{unit volume}}\right) q \cdot dA$$

which, for planar surfaces with constant flow properties, reduces to

$$\text{flux} = \left(\frac{\text{something}}{\text{unit volume}}\right) q \cdot A$$

The "something" here is mass, (X, Y, Z) momentum and energy, which dimensionless and per unit volume are simply R, RU, RV, RW, and E. The area used in the dot product depends on the subroutine being used. Thus

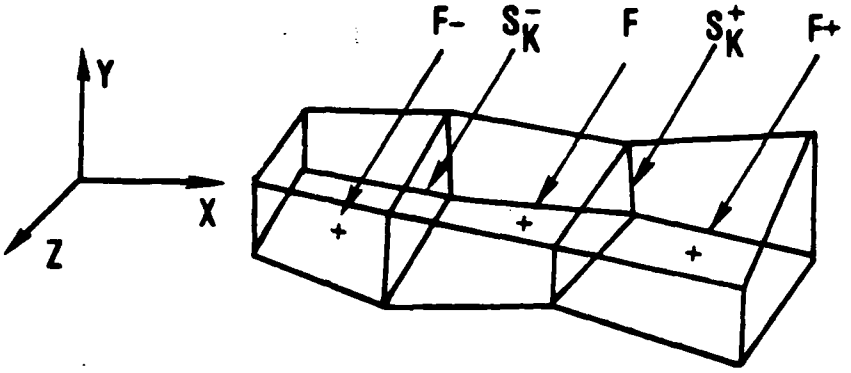
q•A = QDA = (RU*AX1 + RV*AY1 + RW*AZ1)/R in FLUXX,
= (RU*AX2 + RV*AY2 + RW*AZ2)/R in FLUXY, and
= (RU*AX3 + RV*AY3 + RW*AZ3)/R in FLUXZ.

The equations in FLUX then follow from combinations of the "somethings per unit volume" and the QDA terms.

What is not so straightforward in FLUX (and in some of the other subroutines as well) is the way in which the indices (or subscripts) and counters are manipulated. For instance, in FLUXX we find the following indices and DO loop parameters being used: KODE, JJ, JX1, JX2, NEX, NNY, NNZ, JS, JX, JY, JZ, and JT. FLUXY and FLUXZ have similar quantities. Actually, this rather complicated situation is the result of economizing the computations. The solution process involves performing a set of calculations, namely those in equation 11, Ref 1, for every volume element in the computational mesh and at every time step. Within equation 11 we see that there is a need for repeated determinations of F•S, although the subscripts and superscripts appearing in equation 11 suggest that different values of F and S are to be used in the various terms. This is indeed true though with careful study of the overall computational process we find that certain calculations are exactly repeated. For example, at a given time step the quantity $F \cdot S_k^+$ for one volume element is exactly the same (except for a sign) as $F^- \cdot S_k^-$ for a neighboring volume element. (Reference 1 and Section IV of the main paper may help clarify this.) In any case the important idea is that certain types of calculations, namely F•S, are done very many times in this program and some are exactly repeated. The design of the FLUX subroutines is such that the single equations for mass, momentum, and energy flux can be used for every volume, every time step, and each form of F•S in equation 11. The versatility of these subroutines comes from the manipulation of the indices and counters.

A number of subroutines besides FLUX are involved in these manipulations. However, much of this activity does take place in FLUX and a feel for what goes on can be obtained by studying some of the details in the FLUX routines, which we shall now do (on a very limited scale).

FLUX evaluates the quantity F·S. The particular values of F and S used in any given calculation depend on which volume is being considered, which face of the volume is being considered, the time step, and which step of the predictor-corrector technique is being executed. Suppose we fix the volume, face, and time step in question. A representative situation is sketched below.



The volume in question is the middle volume, the face of interest is the shaded one labelled S_k^+. With the time, volume, and face fixed we only need to worry about which step in the predictor-corrector cycle we are in. The two steps in the predictor-corrector cycle are represented by the two equations in equation 11, Ref 1. Looking at these equations we see that we use $F \cdot S_k^+$ in one step and $\overline{F^+} \cdot S_k^+$ in the other. (Don't worry about the bar over F^+ right now.) So, in one step of the predictor-corrector we use the value of F "behind" the face in question; in the other step we use the value of F "ahead." This can be accomplished by using the same equation for F·S and simply changing the subscript of F in the k direction. (Remember, the subscripts of a flow quantity refer to the volume in which it occurs.) If the k direction in this example is in the X direction then we increment JX to go from one side of the face to the other. Such a procedure must be done for all six faces of this volume. The face labelled S_k^- in the sketch would use F^- and F, and so on.

Besides determining F•S at each face, FLUX sums or accumulates the net flux into each volume by incrementing the flux accumulators DR, DRU, DRV, DRW, and DE. However, care must be taken with the signs. For instance, in the sketch the product $F \cdot S_k^+$ might represent the flux <u>out</u> of the middle volume or the flux <u>into</u> the right hand volume. For one volume this number is positive, for the other it is negative. The subscripts and signs must be carefully matched to properly sum or accumulate the fluxes.

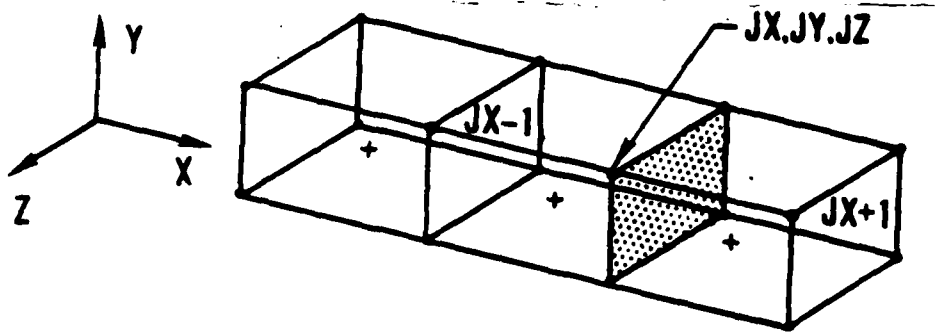
The above, overly long discussion hopefully has established the reasons for the multitude of indices and index operations in the program. What follows is, perhaps, an overly brief summary of how some of the subscripts provide a means to switch from one side of a surface to another and to switch from one volume to another.

Each of FLUXX, FLUXY, and FLUXZ has a dimensioned variable KODE (JJ) whose 16 values are specified in a DATA statement in each subroutine. The value of JJ is established in STEP and is transferred through the labelled COMMON/CONST. JJ tells FLUX which phase of the predictor-corrector technique to compute by picking a value of KODE. KODE, in turn, sets JX1, JY1, or JZ1 depending on which FLUX subroutine we are in. JX1 (or JY1 or JZ1), in turn, sets the value of JT. JT, in turn, sets the value of JS and JS tells FLUX whether to use the flow value on one side or the other of a surface in evaluating F•S and whether to give this quantity a plus or minus sign when it is added to the flux accumulator (schematically represented here by D).

The following chart presents the values of and relations between the various indices as the calculations in FLUX proceed. Also depicted are representative volume elements and reference surfaces to illustrate the predictor-corrector selection and flux summation.

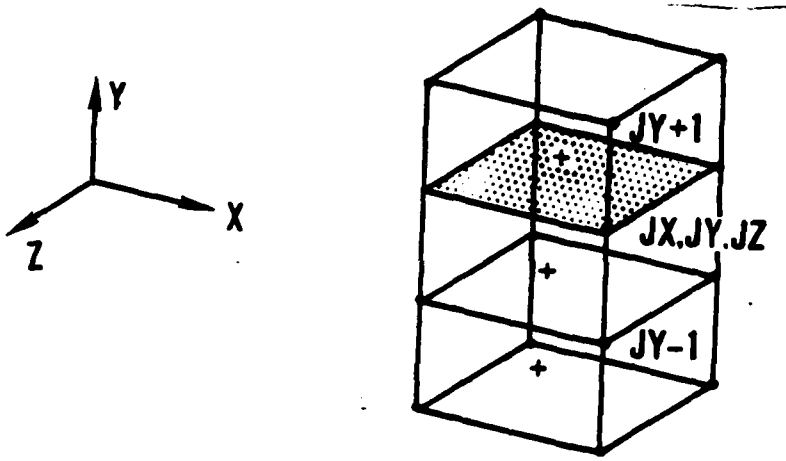
FLUXX

JX1=1 ⇨ JT=2, JS=JX, D(JS)=D(JS)-F(JS-1)+F(JS) ; q upstream of A1
JX1=2 ⇨ JT=3, JS=JX-1, D(JS)=D(JS)-F(JS)+F(JS+1); q downstream of A1



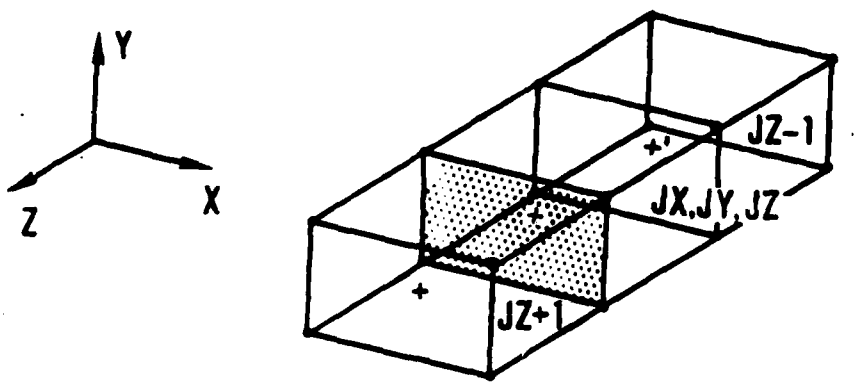
FLUXY

JY1=1 ⇨ JT=2, JS=JY , D(JS)=D(JS)-F(JS-1)+F(JS); q below A2
JY1=2 ⇨ JT=3, JS=JY-1, D(JS)=D(JS)-F(JS)+F(JS+1); q above A2



FLUXZ

JZ1=1 ⇨ JT=2, JS=JZ , D(JS)=D(JS)-F(JS-1)+F(JS); q counterclockwise of A3
JZ1=2 ⇨ JT=3, JS=JZ-1, D(JS)=D(JS)-F(JS)+F(JS+1); q clockwise of A3



A.4 OPERATING INSTRUCTIONS

To run the program, the user must decide on and enter certain information concerning the body, approach flow, and computer sequencing. Let us assume that a run of a single body without the internal clock, restart, or boundary layer options is desired. The user must consider the following items.

Body.

The body presently in the program is a body of revolution previously studied at AEDC (see Fig 8 of Ref 3). If a new body shape is desired the easiest way to input it is through DATA statements in CONE giving the body coordinates (XB,YB). If new coordinates are entered here, certain other changes are necessary also. The DIMENSION statement immediately preceeding DATA XB must accommodate the new body. It would be best to "comment out" the original data in the event a check or reference case is ever needed. The DO loop immediately after DATA YB, which non-dimensionalizes the body coordinates by the body length, should also be adjusted for the new shape.

Grid.

If the body shape entered is axisymmetric and does not have any places on it where the body slope $(dy/dx)_B$ is vertical then the grid schemes described under the section CONE will be satisfactory although they may need some adjustment. For any grid different from 80X32X2 it would be best initially to use the first modification described in CONE.

The grid size must be kept within the limits of the computer. The absolute maximum size using the Eglin CYBER 176 is $600{,}000_8$ ($196{,}608_{10}$) words in large core memory (LCM) and $310{,}000_8$ of regular memory. The storage requirements for this program can be roughly estimated by

STORAGE = 24*XYZ + 21*(XY + XZ + YZ + X + Y + Z)

where (X,Y,Z) stand for (NNX,NNY,NNZ). Almost any grid will be large enough to require LCM. For the LCM option it is usually sufficient to define the labelled COMMON regions AVCTRS, DVRBLS, and VARBLS as LEVEL 2 (thus putting them into LCM) and the rest of the variables will then fit into regular memory. The actual grid size is input simply by changing the PARAMETER statements (all 17 must be changed). The grid scheme used must accommodate the grid size.

If a non-axisymmetric body or a body which has vertical slopes on it is input then a new grid scheme must be written. Note, even a constant diameter circular cylinder with a hemispherical nose has vertical slopes (at X = 0) and must be treated differently. The problem with vertical slopes is that the JX and the JY lines get confused since both will be vertical near the vertical part of the surface. The JY lines must be specialized for these regions. For a non-axisymmetric body it may also be necessary to have variable spacing between JZ lines.

Time.

There are two kinds of time relevant to this program, flow time and CPU time. Flow time is a computed quantity, being the summation of the time increments or steps determined by the subroutine DTCAL. CPU time is the actual time used during the computation process. CPU time is real and depends on the flow time; flow time is an artifact of the calculations and is only realistic in very special situations.

To compute a steady flow, time is advanced until a solution at one time step is sufficiently close to the solution at the previous time step to satisfy the user. The flow time at which this occurs is relatively unimportant. Instead, what is more useful is the number of steps required

to reach this converged solution. These steps are counted by the variable KOUNT. The progress of the solution may be most easily followed by using KOUNT as a timer. There is no direct check on convergence in this program. A user must either know or guess how many steps will be required for convergence. The number of steps used in a given run is specified by the product NB*NT. Many solutions require on the order of 4000 steps to converge and so typical values might by NB = 20, NT = 200 to give 20 printouts every 200 steps for a total of 4000 steps.

The computer time, CPU time, depends on the total number of time steps and the grid size, increasing with both. If the internal computer clock is used, CPU time can be used to control the program. Otherwise it only influences run priority and cost. As examples of CPU requirements, a 30X12X2 grid with 4000 steps requires 400 seconds while a 44X20X7 grid with 4000 steps requires about 80 minutes on the Eglin machine.

Freestream conditions.

The freestream Mach number and direction of the freestream (ALPHA, BETA, and PHI in degrees) must be read in on one of the data cards.

Other parameters.

For no use of restart, internal clock, or boundary layer, set FACT = 0.8, TFLAG = 7.0, and KIN, KOUT, NV, and FACTX all to zero.

Output.

The amount and type of output must be decided upon. Currently printouts of surface pressure coefficients are made NB times, after every NT steps. Also, all the flow variables and grid locations are printed for the last step,

KOUNT = NB*NT. (There is a little problem with this last printout right now; some phantom points also appear in the listing.) The WRITE statements which do all this are in PVAR. PARRAY is also for writing but is currently commented out.

Miscellaneous.

In principal there is no limit on the value of M_∞ used in this program. However, there are practical constraints which limit the useful range to perhaps $.5 < M_\infty < 3$. This should be explored however.

In principal there is no restriction on the body shape which is input but, again, practical limits exist. Any body shape on which there are large separated flow regions will be poorly handled. Separation, at this point in time, cannot be calculated. The grid must accommodate whatever body is used and also have clearly defined volume elements.

SUBROUTINE VOLGRD requires several averages to be formed. These averages involve division by 2, 3, 4, and so on. These divisions never explicitly appear in the program (in order to save computation time). Instead, the divisions are actually absorbed into the time step so that the flow time is off by some constant factor. This has no bearing on the solution except when a truly unsteady flow is to be calculated.

A force and moment calculation program is needed but the boundary layer calculation must first be made operational.

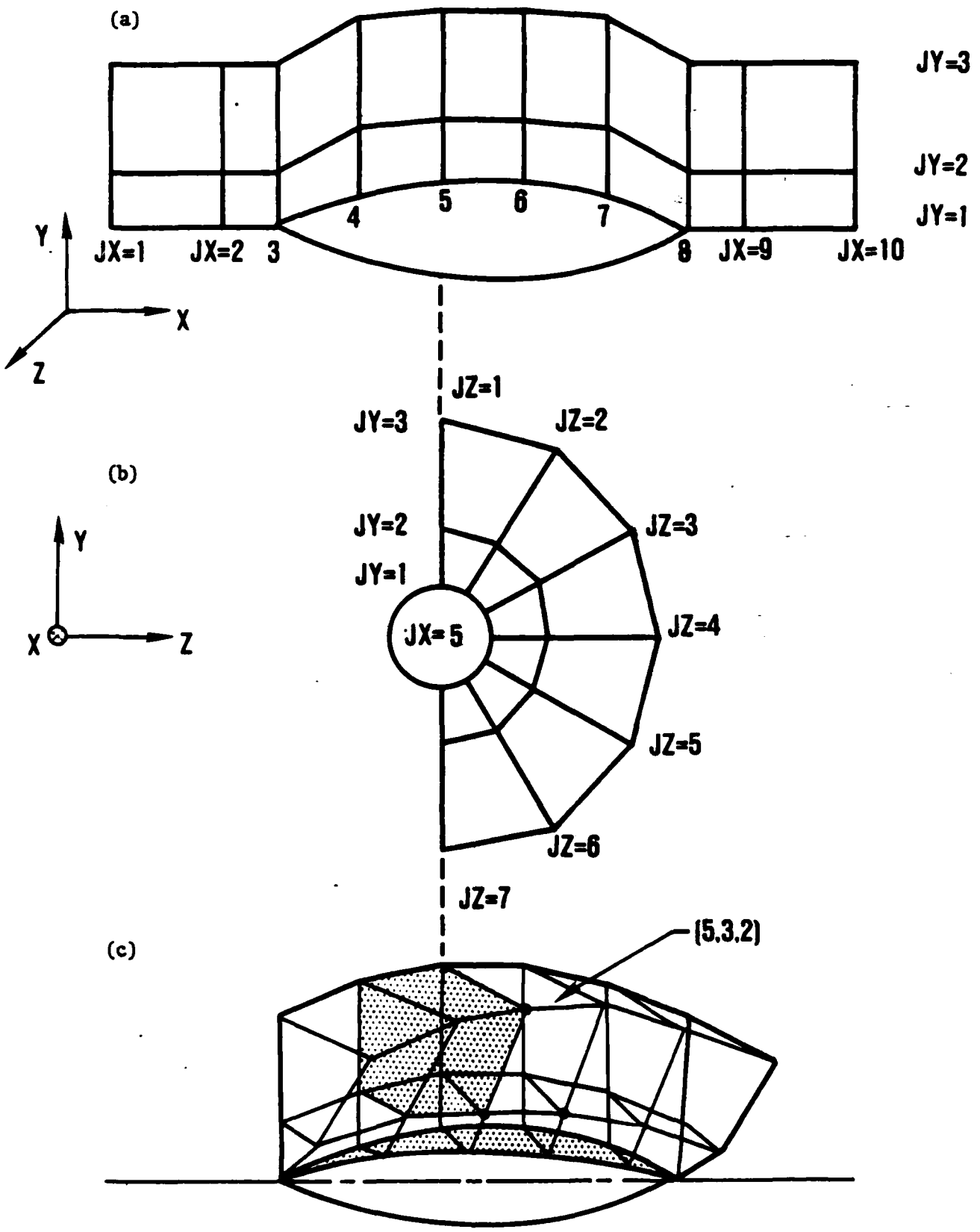


Figure 1. Simple axisymmetric grid

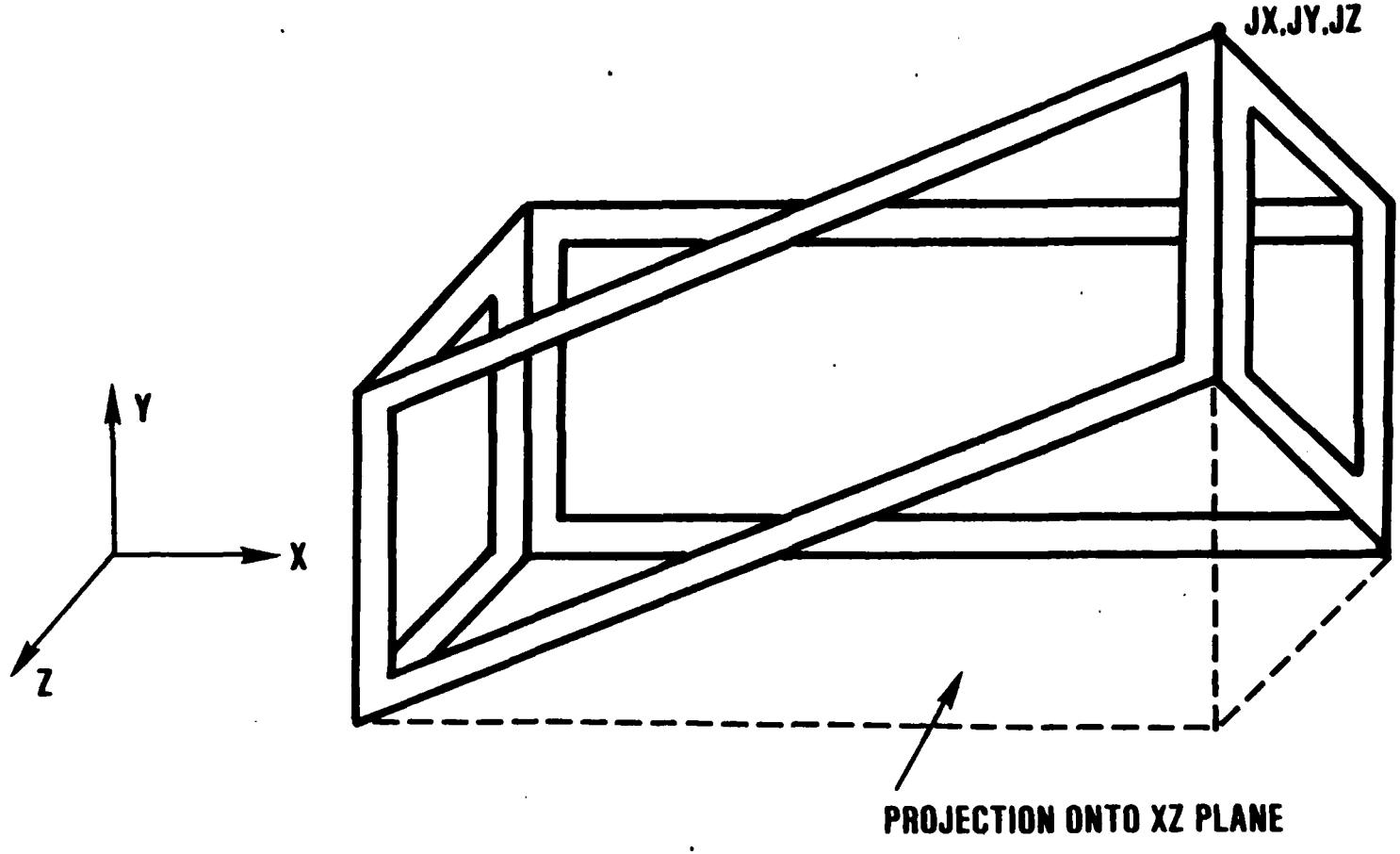


Figure 2. Volume element with non-planar surfaces

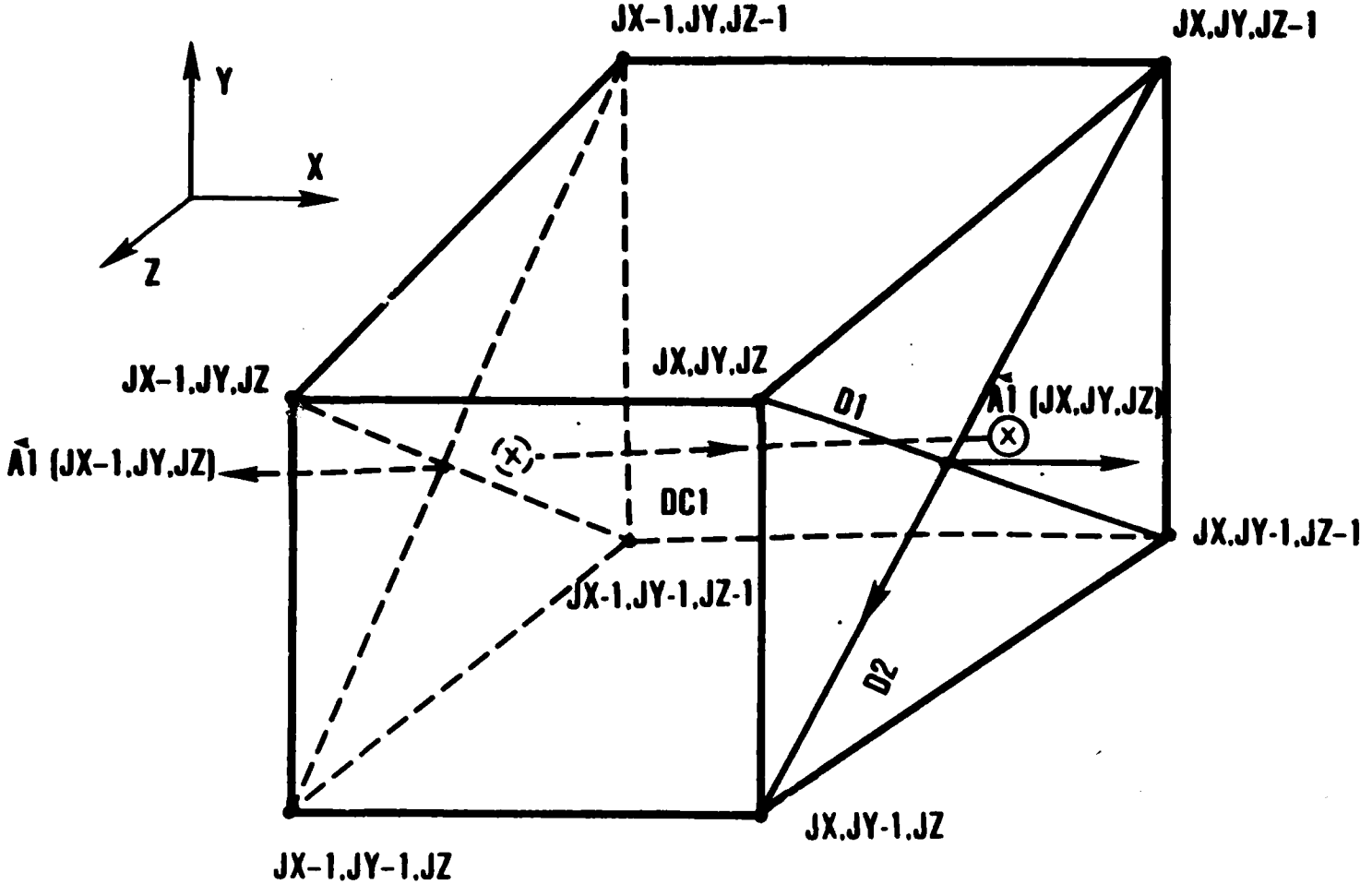


Figure 3. Metrics of a volume element

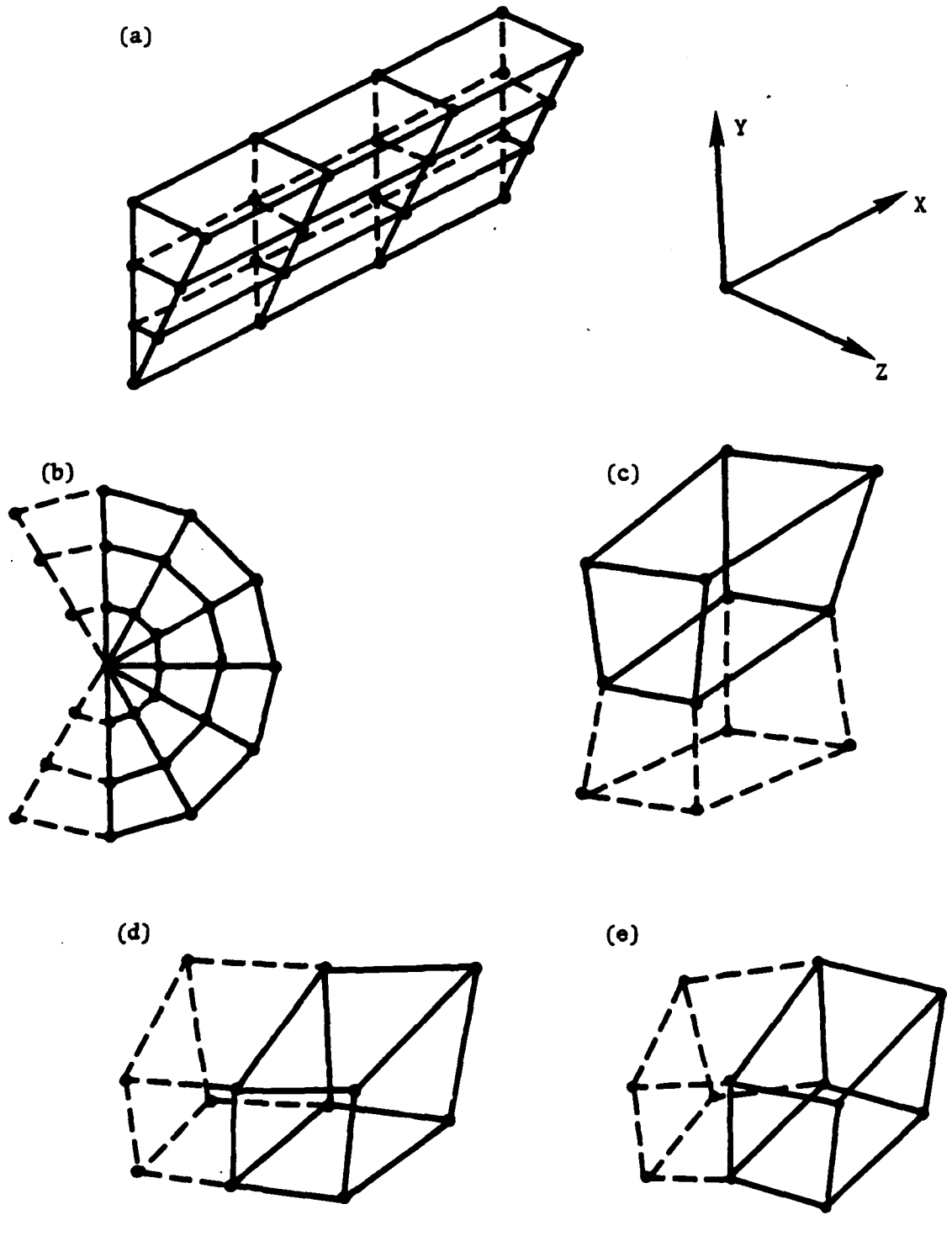


Figure 4. Volume elements and phantom points

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the
SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

DISPLAYING RESULTS FROM NUMERICAL CALCULATIONS USING COMPUTER GRAPHICS

Prepared by:	Dr. Letitia Korbly
Academic Rank:	Assistant Professor
Department and University:	Mathematics Department University of Alabama in Birmingham
Research Location:	Air Force Weapons Lab, Chemical Laser Branch, Computational Group
USAF Research Colleague:	Dr. Nazareno L. Rapagnani
Date:	August 13,1982
Contract No:	F49620-82-C-0035

ACKNOWLEDGEMENTS

I want to thank Larry Rapagnani first of all for making it possible for me to work on this project and for expediting so many details. I want to thank Bob Conley for his close cooperation. Things would have been easier if my work were as clean as his. I want to thank Gary Belie for many conversations, and for his assistance in evaluating my work and this project.

There are many other individuals whom I should thank, who helped me understand things, or helped forward the project, or only shared their time with me. I must here mention Hazel Seely, both for these things and for her help in preparing this report.

Finally I want to thank the Air Force Weapons Lab, whose facilities I've enjoyed, the Office of Scientific Research, who sponsered the work, and Southeastern Center for Electrical Engineering Education and Warren Peele, who funded it.

DISPLAYING RESULTS FROM NUMERICAL CALCULATIONS USING COMPUTER GRAPHICS

by

Letitia Korbly

ABSTRACT

Exploration of specifications for a data base to link results from numerical solutions of partial differential equations with graphics programs which display this information. More generally, two and three dimensional functions defined at a finite number of points are to be graphed so that important features of these functions are easily and accurately grasped.

In preparation, a program linking the output of a two dimensional fluid dynamics calculation with R. W. Conley's Raytrace was written. Refractive index, opacity and color were used to show the differences in density between different nodes in the calculation.

The study indicates that simply using a standard data base will not solve the problems connected with linking three dimensional digital data with graphics programs. Optical properties must be chosen to highlight functional values. Color and opacity are the easiest qualities to interpret.

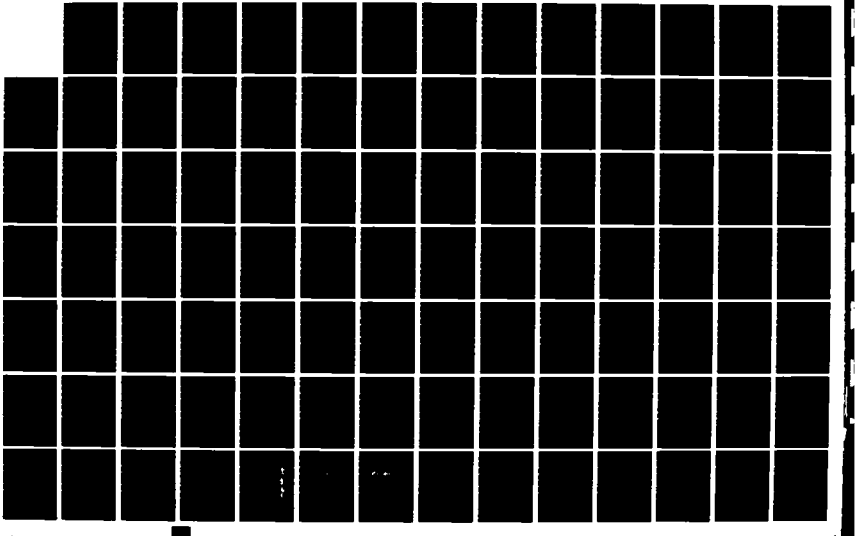
Normalization is crucial to effective display. Normalization matches the range of the data to the range of the display so the data fits the display range.

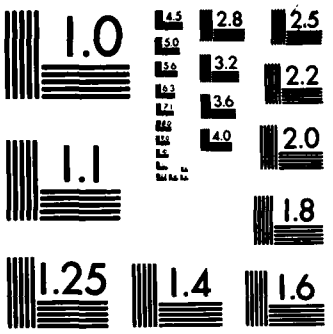
While using refractive index to display density offers the possibility of evaluating the analyses directly at some time, it is presently only confusing.

Suggestions are made for further research.

AD-A130 770 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM (1982) RESEARCH REPORTS VOLUME... (U) SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION INC S... 2/11

UNCLASSIFIED W D PEELE ET AL. OCT 82 AFOSR-TR-83-0614 F/G 5/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

DISPLAYING RESULTS FROM NUMERICAL CALCULATIONS USING COMPUTER GRAPHICS

by

Letitia Korbly

ABSTRACT

Exploration of specifications for a data base to link results from numerical solutions of partial differential equations with graphics programs which display this information. More generally, two and three dimensional functions defined at a finite number of points are to be graphed so that important features of these functions are easily and accurately grasped.

In preparation, a program linking the output of a two dimensional fluid dynamics calculation with R. W. Conley's Raytrace was written. Refractive index, opacity and color were used to show the differences in density between different nodes in the calculation.

The study indicates that simply using a standard data base will not solve the problems connected with linking three dimensional digital data with graphics programs. Optical properties must be chosen to highlight functional values. Color and opacity are the easiest qualities to interpret.

Normalization is crucial to effective display. Normalization matches the range of the data to the range of the display so the data fits the display range.

While using refractive index to display density offers the possibility of evaluating the analyses directly at some time, it is presently only confusing.

Suggestions are made for further research.

I. INTRODUCTION:

When partial differential equations in two or three dimensions are solved numerically, a functional value is generated at each node or element in the mesh. These nodes fill the area or volume in which the calculations are done.

In two dimensions a small problem may involve 50 by 24 or 1200 locations. In a full three dimensional calculation, there may be upwards of ten to fifteen thousand nodes involved. At every time step information is calculated for every point. Fluid dynamics calculations typically involve several functions evaluated at each point at each time step.

Traditionally graphing has been used to handle large amounts of numerical data. Many techniques are available if the data lie on a line.

For data in a plane, a recent study by R. G. Belie indicates that it is easier to interpret carpet plots than other forms of presentation. This technique does not easily generalize to three dimensions. Color-coded plots provide the next easiest form to interpret. Color is not so difficult to lift to three dimensions.

Contour plots are adequate for display of results in structural mechanics since maximum and minimum values, generally lie on the surface. Contour plots generally require as much computer time as color plots.

For three dimensional functions the contour picture is inadequate. In a three dimensional fluid dynamics calculation, it is important to see what is going on deep inside. Conventional graphics programs cannot show this.

Color graphics, with the possibility of tracing light rays through transparent layers, offers a variety of possible answers to the problem. Changes in color, opacity, and brightness can be used to show the differences in the calculated values. It is necessary to find or develop programs which utilize these possibilities.

To see deep into a fluid and determine differences in density and velocity there, something must be either transparent or stripped away. Stripping almost implies contouring, a difficult job and one that is slow on a computer. This leaves the possibility of treating the layers

as if they were transparent or translucent. Two possibilities for handling this have been advanced, ray tracing and Fourier transform techniques. It is not clear how to implement Fourier transform techniques in an area with irregular boundaries.

It is also important in these calculations to see the velocities. This information can be well presented by particle tracing. Frequently this information is presented by taking a section through the volume and showing the component of the velocity which lies in the plane as an arrow pointing in the same direction as the component of the velocity lying in that plane. The length of the arrow is proportional to the magnitude of the component of the velocity lying in the plane. This can be confusing, particularly if the total velocities are highly varied.

Right now it is still not routine to show density differences meaningfully. Since density is linked with refractive index in the physical world, it was hoped this would provide a right way, or several right ways, to display density.

There is another important issue: verification of the computer codes which produce the data. In the past, few full three dimensional calculations were done because of the amount and expense of computer time involved. It is difficult to verify and evaluate the results of these calculations. Refractive index is easily measured in a moving fluid without disrupting the flow. This suggests the possibility of direct evaluation of these programs against measurable physical data. Preliminary study indicates, however, that this is further from realization than simple display of data.

It is probable that initially these programs will be expensive to implement and run. As more use occurs, optimization should reduce costs.

My background includes a specialty in partial differential equations, and applied mathematics. I have had experience in fluid dynamics, at Space Sciences Laboratory, Marshall Space Flight Center, Huntsville, Alabama, and in the development of graphics programs, at David Taylor Naval Ship Research and Development Center in Carderock, Maryland. My interest in these problems is high.

II. OBJECTIVES

The main objective is the development of standard tools and packages to display results from computing solutions to partial differential equations, especially solutions to three dimensional fluid dynamics problems. Secondarily, it is hoped that such techniques may be used to display more general two and three dimensional functions defined at a finite number of points. This data might include information about the levels of geologic strata deep in the earth, currents in the oceans and atmosphere, etc.

A lower priority objective was the evaluation of analytic programs. It is possible to measure refractive index in a moving fluid without disturbing the fluid. Such a measurement, compared with a frame which displays density using refractive index, could be used for an evaluation of the analytic program since the link between density and refractive index is linear through the Gladstone-Dale constant.

Clear understanding of the problems involved is necessary to establish such a general link between programs that solve partial differential equations numerically and graphics programs. Raytrace is a three dimensional graphics program which is not in production form, but which has the important capacity to see through transparent and semitransparent objects, taking into account important optical qualities through actually following light rays through the objects and medium. Though not in production form, it is certainly clean enough to make an attempt to use it to display such data.

The immediate goal was to implement Raytrace as a graphics output program for a two dimensional density function calculated by a fluid dynamics program. The long term goal is to use Raytrace, or other programs routinely to display output from numerical solutions of partial differential equations in three dimensions. The system of equations of greatest interest is that of fluid dynamics, which is both three dimensional and time dependent.

Full three dimensional graphics programs are needed to show what is happening deep in the fluid. These programs almost necessarily do ray tracing although programs utilizing a Fourier transform have been discussed.

Production programs which do ray tracing are very specialized. Such programs are commonly used in optics, but are generally unsuitable for displaying the output of a numerical process.

R. W. Conley's Raytrace is a full three dimensional graphics program almost in production form. It is presently the only practical program for display of optical qualities in translucent materials like color, opacity, and refractive index of solids, and transmittance, albedo, and mirror properties of their surfaces as well as other optical properties which were not exploited in this study.

It does not, however, accept data in the form supplied by most analytic programs. To utilize Raytrace, it is necessary to write programs which link the graphics package with the existing data.

This problem is characteristic of attempts to utilize graphics. There is no standard hardware or software; each data source must be individually linked with each graphics package, and each graphics package must be linked individually with each output device.

If a standard data base were established, writers of both kinds of codes would be pushed toward adherence to that standard. This will lead to development of graphics programs which can utilize standardized output from the analysis packages, or to graphics programs which can be easily linked to the data base.

III. PRELIMINARY STUDY

In preparation for the design of this data base, a program to link output from a simple two dimensional fluid dynamics to a three dimensional graphics package was written. Conventional graphics programs have not been able to show the results of three dimensional fluid flow calculations satisfactorily. It is necessary to implement a three dimensional graphics program that can see through surfaces.

Raytrace was chosen because it is the only program which can show the interiors of transparent, partly transparent and translucent objects like fluids.

IV. LINKING NUMERICAL OUTPUT TO GRAPHICS

Raytrace has great potential for showing the results of these calculations because it traces light rays. More precisely, Raytrace calculates the light arriving at each point (pixel) on the screen from each source in the chosen geometric configuration. It adds the direct, refracted, reflected, and diffused light arriving at the point from each light source as well as calculating and allowing for light absorbed or emitted by translucent solids.

This means at least one and usually several calculations must be done for each pixel on the screen in a small implementation, and many calculations for the implementation envisioned in displaying the results of three dimensional fluid dynamics calculations. In doing calculations with many objects, Raytrace uses very large amounts of computer time and CPU storage, as well as storage of the files that go to the display terminals and other facilities.

Several Fortran programs were written to provide data in an appropriate format for Raytrace. This data is then used to generate a red-green-blue file. This step takes very large amounts of computer time and memory. The increase in time seems proportionate to the number of objects, and is dependent on the complexity of the optical properties of the objects.

The red-green-blue file is then submitted to drivers which produce data formatted for the display devices, the AED-512 color graphics terminal and the FR80 film recorder. These files are finally used to drive the output devices, which give pictures on the screen or film to be shown.

As originally implemented, a single frame with roughly twenty objects took about 600 seconds on the Cray-1S and about twelve hours on a Vax. With cubes arranged in a six by thirteen grid, the Cray took more than 500 seconds to calculate the red-green-blue file for a frame which used color to display the density differences, almost the same for a frame which used opacity, and upwards of forty minutes cpu time for each frame which used refractive index to display density differences.

It is clear that to run many frames a very large computer is necessary. Present codes and computers do not easily handle programs of this size and complexity. The amount of time used in running the three dimensional fluid dynamics codes is much larger than this. Rapagnani and Belie estimate that the cost of displaying output can reasonably be one-tenth the cost of running the code.

V. THE COMPUTER SYSTEM

The computer system used was the Cray 1S at the Air Force Weapons Lab at Kirkland Air Force Base at Albuquerque, NM. It is frontended by a CDC-750 and has an IBM mass storage facility. Long term files storage is accessed through the CDC-750. Continuing problems with the computers plagued and retarded the project.

VI. RAYTRACE DATA AND DISPLAYS

Raytrace needs information on the color and brightness of the background; it needs at least one light source. The structures of Raytrace are spheres, planes, which are simply rectangular surfaces which reflect or block light, lenses, or right circular cylinders, and polyhedra. These structures have geometric properties that completely define them in space, and optical properties that allow the viewer to perceive them and objects reflected in them and objects visible through them.

Raytrace does not generate polyhedra automatically. Polyhedra are generated by giving their bounding surfaces, which must be convex polygons. Polygons are defined by giving their nodes. It is the responsibility of those who submit the data to determine that the polygons are in fact convex, that the polyhedron is in fact closed, and that the polygons do not intersect other objects in the space.

The optical properties primarily exploited were color, opacity, or coefficient of transmittance of light through a solid, and the refractive index.

An optical property was linked to the density of the fluid. When color and opacity were the optical properties used, the density was multiplied by a constant to normalize it and this value assigned to the optical property. Refractive index was calculated as 1.0 plus the density divided by the Gladstone-dale constant.

It required some experimentation to choose values of the remaining optical properties which would make the presentation clear. It is possible to assign albedo, transmittance, and mirror qualities to the surfaces of the polyhedra. Albedo is the coefficient of diffuse reflected light. The two dimensional pilot study showed that albedo and the coefficient of mirror reflection should be set to zero so that the outlines of the cubes would not dominate the visual presentation. Transmittance, the percentage of light passing through the surface, was set at one hundred percent for the same reason. The net result was to make the boundries of the polyhedra invisible.

A cube was generated at each node to represent that portion of the fluid. Each cube then represented a location. Optical properties of the cubes varied proportionately with the density of the fluid. When one property was tied to the density, the others were held fixed. Some future experiment should vary these parameters simultaneously.

The polyhedra were originally set up tightly packed. Later in the study it appeared that using the same surfaces on adjacent polyhedra might confuse Raytrace, so the polyhedra were slightly shrunk to separate the bounding surfaces.

The optical properties were assigned using data statements. This was not initially done, but several times current files were lost due to hardware and software problems. Sometimes the calculations proceeded using optical properties different from those planned. Since there was at least a full day between turnarounds, it was some time before these errors were caught.

The cubes were generated by the program, and for those which used color or opacity to display light, the color or opacity were calculated and assigned at this time.

To perceive attenuation or refraction of light, some object which the eye expects to see in a given place must be obscured by the attenuation of light or displaced by the refraction of light.

When refractive index was the property used to display density differences, objects which would appear displaced by the refraction of light were generated by the program. These objects were simply square flat black squares placed immediately below the centers of the cubes.

After the first frames using data actually generated by a fluid dynamics program were seen, it became clear that normalization would be required to see the results. It was not possible at that time to redo the code, so the constant that normalizes the data was chosen so that the maximum value of the property which displays density differences would be the maximum value expected by Raytrace.

VIII. RESULTS

Data cannot be presented so it can be easily and accurately grasped by simply matching data bases. It will be necessary to decide on a method of display, the center of interest, the point of view, and so forth. Some intelligent linkage with the graphics program will be necessary. The following paragraphs document this conclusion.

The frames created showed that color and density are most easily perceived of the optical qualities studied. The eye perceives very subtle color differences. This became clear when a series of unnormalized frames was developed in the last week of the study. The differences displayed were very small.

These small differences were easily seen in the frame which used color; they were harder to detect in the frame which used opacity. On the slide which used refractive index showed only through the presence of extraneous marks on a few cubes.

Both color and opacity offered results which were easy to interpret. Color seemed not only more sensitive to differences, but easier to interpret. It was still clear which cubes are more and less transparent. This is expected to hold in three dimensions.

Differences in refractive index are harder to perceive. If the differences in refractive index are large, the resulting optical illusions are too confusing to be processed automatically by the brain. Some frames could not be interpreted. Refractive index is still too confusing for the display of results. In addition, limitations on machine time and resolution of the terminals show more here than with color and opacity.

As opposed to refractive index, both color and opacity give frames most easily viewed head on. The light must be oblique to see differences in refractive index. It will however, be necessary to see several different views when finding the most highly colored or opaque areas within a three dimensional object, even if color and/or opacity are used to display the function. The total volume may be rotated while stepping through time if the calculation is time dependent.

In doing the preliminary calculations, several frames were generated which had insufficient contrast. This made clear that it will be very important to normalize the data which is being presented. Any routine which attempts to take the data from its generator to a graphics display will be extremely inefficient if it does not normalize the data.

To display the entire twenty-four by fifty data base took far more computer time than was available, so a six by thirteen section of great interest was displayed. Even so, computer time on the Cray-1S ran from five hundred seconds for frames using color or opacity to show density to twenty-four thousand seconds to prepare the same data using refractive index. Two frames using refractive index ran out of time after three thousand seconds. These frames were similiar to the others which used refractive index, except that the point of view was moved. These programs were not rerun, although it is probable that they would have not taken much more cpu time.

Raytrace takes too long to display the data from even moderately large two dimensional calculations because the speed at which it runs depends on the number of objects. Raytrace might be speeded up by tagging surfaces which are shared by adjacent objects so that a ray leaving an object will automatically enter the next object.

IX. SUGGESTIONS FOR FURTHER WORK

To display the results of fluid dynamics calculations to both specialists and nonspecialists quickly and clearly more work is necessary.

Raytrace should be tested using small three dimensional data bases contrived so that known objects are inside. This would make it much easier to evaluate schemes to display the data. After this, actual results from numerical solutions of partial differential equations can used to test the schemes.

It is also recommended that a program be written to generate data for Raytrace in these applications. It should be self documenting and capable of implementation in both batch and interactive mode. The program should arrange the data set up so that it is easy to check and change the optical property or properties which display the data, the portion of the data to be presented, and the normalization and rescaling of the data.

Some normalization routine is necessary to rationally show the results, and this should be incorporated in any program which links analytic data with graphical display.

The first recommendations depend on optimizing Raytrace. The runtime for Raytrace corresponded to the number of objects and the optical complexity of these objects., A version of Raytrace whose data structure is suited to keeping track of surfaces shared between two adjacent polyhedra should speed calculations significently.

The data considered here is on a grid. All data produced by numerical solution of partial differential equations is either on a grid, like finite difference results, or has space filling objects already generated, like the elements in finite elements. In the first case it is very easy to generate rectangular polyhedra to fill the space by centering one at each node.

Output from numerical solution of partial differential equations is data appropriately presented as filling the area or volume of interest with non-overlaping objects. This space can be viewed is filled with packed forms. All data returned by programs which solve partial differential equations is in this form or can easily be put in this form. For

this application then, Raytrace will be most efficient if the data structures tag surfaces so that as the ray leaves one object through a surface, the data structure is tagged so it enters the the adjacent object which shares this common surface without a search.

Since Raytrace is run on vector machines, optimization gains can be made by vectorizing the code.

I propose to carry through the first three proposals, using data from N. L. Rapagnani's fluid dynamics calculations as part of the test. This proposal will be feasible only with small data bases unless Raytrace is optimized. I also propose that a contract be established with R. W. Conley to write a version of Raytrace optimized for presentation of results from numerical solution of partial differential equations.

The proposed study will concentrate on display of results using color, opacity, and combinations of the two. It will also explore options of looking at three dimensional subvolumes of various types.

After the first study, a standard data base should be developed. While far from a complete answer to the problems involved, this will facilitate changing among graphics programs to display output since authors of both codes will be pressed towards adherence to the standard.

Such a data base should retain the location of the node, the quantity or quantities to be displayed, and elements in which the node occurs if it is a finite element code. Preserving the elements saves regenerating them for Raytrace. Since it will be useful to sort on all these objects, the data base must make provision for this.

REFERENCES

1. R. G. Belie, "Color Computer Graphics in the Engineering Workplace," Second International Computer Engineering Conference, San Diego, CA, August, 1982.

2. R. G. Belie and N. L. Rapagnani, "Color Computer Graphics," Areonautics and Astronautics, June, 1981.

3. L. A. Feldman and N. L. Rapagnani, "Computer Graphics for Scientific Applications," AFWL Report AFWL-TR-80-138, February 1982.

There is no published documentation of Raytrace. It has produced a movie shown at SIGGRAPH 81 during Science Night Films and Video Shows

R. W. Conley, "Refractions," Computer Graphics, Vol. 15, No. 9, p. 315, December, 1981.

and during the following talk:

F. C. Crow and M. W. Howard, "A Frame Buffer System with Enhanced Functionality," Computer Graphics, Vol. 15, No. 3, pp. 63-9, August 1981.

This movie has been shown twice on television:

First Computers Are People Too, a Walt Disney Production, June 6,1982.

Walter Cronkite's Universe, July 6, 1982.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

MODIFICATION OF CURRENT FEEDBACK STRATEGIES: A TEXT SYNTHESIS APPROACH

Prepared by:	Dr. Philip Langer
Academic Rank:	Associate Professor
Department and University:	Educational-Psychological Studies University of Colorado
Research Location:	Air Force Human Resources Laboratory Lowry Air Force Base, Colorado
USAF Research Contact:	Dr. Robert H. Summers
Date:	July 26, 1982
Contract No:	F49620-82-C-0035

Acknowledgement

I would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research, and the Southeastern Center for Electrical Engineering Education for providing me with an invaluable summer experience at the Air Force Human Resources Laboratory, Lowry Air Force Base, Colorado.

I would especially like to acknowledge the invaluable discussions with Dr. Edgar A. Smith and Captain Richard T. Dineen of the Laboratory, and MSgt. Joseph E. Burner (3450 TCHTG/TTMYA). And finally, I should like to express my gratitude to Dr. Robert H. Summers, without whose stimulating knowledge and guidance this project would not have been completed.

MODIFICATION OF CURRENT FEEDBACK STRATEGIES: A TEXT SYNTHESIS APPROACH

By

Dr. Philip Langer

ABSTRACT

Training with respect to maintenace of sophisticated weaponry is becoming an increasingly more complex Air Force problem. Curriculum sequencing and content seems likely to be little impacted by recent findings in cognitive psychology, and hence feedback becomes an increasingly more significant adjunct aid to instruction. The behaviorist position views feedback as an incentive, assuming learning consists of a strengthening or weakening of stimulus-response associations. The cognitive position argues the critical impact of feedback is one of information, and assumes that the organization and utilization of content is dependent on individual strategies. The behaviorist model appears adequate for hierarchial or fixed sequence content, while the cognitive postition appears more tenable for problem solving and higher order learning. This project proposes the exploration of the parameters of the feedback mechanism from a cognitive viewpoint, utilizing a text synthesis method. Through analyses of learner outcomes, including errors, a model may be developed which will permit matching of the most appropriate feedback strategy to the learner's cognitive processing. In effect, feedback will be "titrated" as to amount and frequency based on the learner's response.

I. INTRODUCTION

The levels of technological sophistication of Air Force weaponry, as well as support and maintenance equipment, appear to be increasing at exponential rates. Since the readiness of this technology is directly related to military effectiveness, training problems relative to maintenance and support are critical.

Air Training Command is under pressure to develop instructional programs commensurate with the increasingly more complex technology, both currently employed and under development. Moreover trainee abilities have not necessarily kept pace with these learning demands. Unlike any other educational institution, the military is faced with an enormous diversity in both types and levels of courses and ranges of student abilities. Probably the effectiveness of the traditional classroom format has reached its limits, and technology (e.g. computers) will play an increasing, if not paradoxical, dual role both as the problem (weaponry) and the solution (instruction).

Unfortunately, the implementation of technically assisted instruction has not been all that effective. A major problem is that curriculum development as utilized in technology has not taken into account recent advances in cognitive psychology. Generally the assumption has been that computers can initiate and direct learning by providing a sequentially detailed content based on a rational strategy model. Even the so-called tutorial languages follow this premise.

An alternative hypothesis would suggest that the programs should take into account the simple fact that people think; and that trainees at the age levels found in the training courses have developed quite functional and complex cognitive repertoires. Hence instruction, particularly at the more complex levels, might well be considered as learner initiated and directed. Thus programming becomes responsive to learner strategies, and utilizes learner resources already present.

Compounding and confounding our research problem is that the use of instructional technology stresses a 1:1 relationship between machine and learner. Hence, programmatic weaknesses cannot be smoothed over by averaging across classroom differences. To compensate for this problem, elaborate feedback programs have been initiated to meet and resolve learner strengths and weaknesses.

These are programmatically linear models, which as we will show, are not tenable in the light of contemporary psychology. Hence, individualization remains defined at the most primitive functional level (i.e., one student per console). Changes in feedback strategies are needed to meet increasingly more complex individualized instructional environments.

II. OBJECTIVES OF THE RESEARCH EFFORT

The objectives of the project were to (a) describe current research and theoretical findings which demonstrate the need for new feedback strategies, and (b) outline an initial research effort to develop these strategies. Feedback strategies will be analyzed using the text synthesis method, in which meaning must

be constructed from a random presentation of discourse sentences.

III. CURRENT FINDINGS

This paper deals with a number of critical issues concerning feedback in instruction, and proposes a text synthesis research paradigm to investigate these variables. The assumption is that feedback, a significant adjunct aid in instruction, is currently by and large ineffectual except in many simple instructional contexts. Section 1 deals with general issues in education, Section 2 presents a review of feedback literature, and Section 3 reviews scrambled sequence literature relevant to the research paradigms.

SECTION 1

Educational specialists are currently under pressure to demonstrate an increased response to demands for individualized instruction. Such learning approaches are generally defined as systems in which the learning environment is organized and maintained from feedback derived uniquely from each individual student (Langer, Note 1). This stress on individualization reflects increased emphasis on mastery learning (Bloom, 1968), increasing utilization of cost-effective technology, and special needs of students (Gage & Berliner, 1979). The need for effective feedback is obviously critical.

In terms of instructional design, the dominating patterns have included variants of a multi-track, multi-modal model, with considerable usage of programmed learning and computer assistance (Langer, Note 1). All reflect a powerful behavioral bias. However, even with marked trends toward decreasing unit cost for

the hardware and increasingly more sophisticated software, the multi-track approach represents a very limited gain over the single track curriculum. The reason is that under present circumstances the same primitive methods of curriculum construction (and feedback) prevail. As a consequence, while the management delivery components (e.g. technology) are becoming rather sophisticated, models of the psychological processes underlying content orgnization have lagged considerably with respect to systematic development. What changes have occurred, have followed directly from Skinner's (1953) overall view of human learning.

The behavioral position regarding all learning is that of associations formed between stimuli and responses. Learning therefore consists of gradually strengthening or weakening associations as a function of the consequences of the response (Thorndike, 1913). Maltzman's (1955) paper on concept assumption and Bourne and Restle's (1959) work on "cue validity" are all derivatives of this basic postulate. Feedback under these circumstances becomes a mechanistic process of confirming-disconfirming, modifying thought by conditioning or suppressing associations.

Instructionally this model assumes that concepts are arranged in some hierarchial sequence, generally from simple to complex (Gagne, 1977). Such an arrangement, designed to satisfy behaviorally defined objectives, meets the criterion for programmed instruction independent of format (Markle, 1973). Regardless of the curriculum typologies underlying the specific

content (i.e. world-related, concept-related, inquiry-related, learning-theory related, and/or utilization-related), the concept logic tends to dominate (Posner & Strike, Note 2). This means that the sequencing is generally designed to meet subject-matter goals first.

This does not constitute much of a problem in hierarchically derived content such as mathematics or where the objective of the course is to teach a set of predetermined procedures, but these represent a rather limited range of educational experiences. To utilize this approach in other types of learning situations, the developer must accept the operant assumption that environmental manipulations can overcome learner tendencies to significantly transform or modify the content beyond presumed limits.

For example, most variants of the Keller (1969) Personalized Instructional System utilize a single standard text (Robin, 1976). Obviously, the usual text is hardly "programmed" in any sense of the word. The management system therefore compensates for text deficiencies by organizing the content so that the timing for appropriate contingencies is met (Johnson & Ruskin, 1977). Again, the underlying assumption is that learning can be predictively determined through rationally ordered environmental manipulations, operationally defined as a set of direct instructional guides.

However, the prevailing assumption of direct causality between an instructional stimulus (e.g., content, feedback) and subsequent cognitive processing is generally unproven (Winne, 1982; Bilodeau, 1969). Validation of a process based on the

product (i.e., a response) is logically weak, since one has to prove there was no other way the student could have derived the response (e.g., alternative strategies). From the behaviorist viewpoint the assumption of a conditioning paradigm with deterministic transformations (e.g., generalization and discrimination gradients) is absolutely critical.

Unfortunately, the phenomenon of subjective organization including the use of strategies in recall is well established (e.g., Anderson, 1980). Indeed, as early as 1932 Bartlett postulated the principle of constructive memory. Other types of memory organizations have been subsequently derived in addition to subjective; i.e., categorical and associative. A number of measures of subjective organization have been posited (e.g., Tulving, 1962; Shuell, 1969; Pellegrino, 1972). Cognitive strategies, as we shall use the term, involve an active serial processing based on feature relevance (Rosch & Mervis, 1975) and rule learning (Bourne, 1970).

Early findings clearly indicated that under free recall conditions subjective order did not necessarily coincide with the original (input) order. Starting with Bousfield's (1953) findings, subjective organization appears to have been considered either a characteristic or explanation for memory; under either explanatory construct it is an empirically verifiable phenomenon.

Among the more critical variables affecting subjective oganization is order of input. Subjective organization is higher under constant rather than variable order of input (Pellegrino & Battig, 1974; Wood, 1972). Their findings suggested to us that

the usual text analysis approach, with processing order imposed by the discourse logic, might be confounding. In our research we intend to explore the input variable by randomly generating the content (i.e. sentences) remaining after each selection, as compared to maintaining order. In addition, subjective organization tends to increase with increased outputs (Postman, 1972), and blocked presentation also increases subjective organization (Buschke, 1976).

The meaningfulness of the material appears to influence recall, although the findings have not established a clear pattern (e.g., Glidden & Roemer, 1974; Gorfein & Blair, 1971). Our text synthesis approach we feel will enable us to explore more fully this issue since we will explore both meaningfulness in terms of original content as well as that suggested by changes in concept strategies as reconstruction develops. In one sense the sequential reconstruction of the discourse may have implications for the part-to-whole and whole-to-part transfer problem (Postman, 1972), and should provide an additional window to the structural propositions of individual semantic memories.

Finally, subjective organization appears to be related to a more general memory ability. Subjective organization appears to correlate with free recall, serial learning, and associative learning of word lists (e.g., Earnhard, 1970). This is a critical aspect of our project, since our assumption is that we are dealing with a stable (if not unique) set of abilities, rather than more situational (and temporary) states.

Putting it another way, while subjective organization may

vary from individual to individual, there is stability within the individual. Thus Kozminsky, Kintsch, Coren, & Bourne (Note 3) found at least four different strategies in a moderately complex decision-making problem, while Sternberg and Ketron (1982) found that untrained subjects in an analogy solving problem tended to utilize reliably effective self-generated strateges. As we shall note in our discussion of feedback, these individual differences lead to problems involving linear feedback models, since the developer's assumptions about what is needed may differ markedly from those of the learner at any point.

Although organization is assessed at recall, the differential processing is likely to have occurred during input (Mandler, 1972). Clark and Haviland (1975) point out that inferences are made during acquisition, resulting in an ever-changing memory structure. This necessitates recognition of memory as a developmental phenomenon. Thus, going back to our discussion of the interaction between content and learner strategies, we find for example the intra-individual strategies undergo modification from more simple to complex (Levine, 1975).

The thesis can be made that recent advances in cognition have invalidated many earlier behavioral assumptions about learning. To the curriculum developer, particularly in learning situations depending on a heavy feedback component, the issues have grown exceedingly more complex. We have argued that behavioral management systems have followed Skinner's dictum of small controlled steps to reduce error, and as a consequence feedback has been dictated by the assumption that the strategy

minimizes significant individual transformations of meaning.

The paradox is that while modern cognitive theory can be described accurately as information-processing (Bourne, Dominowski, & Loftus, 1979), the cognitive theorists have yet to systematically address problems of content and sequencing, especially in higher-order learning environments. The issue of feedback is scarcely addressed in instructional terms (Clark, Note 4). Behavioral approaches, on the other hand, have had a great deal to say about content and sequencing as well as feedback, but have generally ignored learner transformations. This paper will explore the impact of more recent cognitive findings on feedback, a critical adjunct aid.

Section 2

Feedback cannot be easily defined, primarily because a number of theoretical positions are embedded in the concept today. Since the instructional implications of feedback may be considered originating essentially from the work of Thorndike and The Law of Effect (Thorndike, 1932), behaviorists have defined it as the consequences of a given response and made it operationally part of the conditioning paradigm (Anderson & Faust, 1973). Thus, one could speak of reward and punishment as confirming and disconfirming consequences respectively, and feedback overall in terms of an incentive model (Buss, Braden, Orgel, & Buss, 1956). The feedback term "knowledge of results" (KR) can easily be utilized by behaviorists, if the confirmation/disconfirmation produces concomitant changes in probability of response through

conditioning. Instructionally, more informative KR models can be maintained within the conditioning paradigm (Anderson & Faust, 1973), since the behavioral position does not deal directly with the input of information on incentive, per se. The cognitive position (e.g., Atkinson & Wickens, 1971) assumed that information contributes directly to the immediacy and direction of change.

Typically, the behavioral experimental paradigm involves minimal informational cues aside from the rewards and punishments utilized. Currently it is assumed that the effectiveness of feedback under these conditions is a function of a number of variables, some of which will be discussed later. However, recognizing the necessity of immediate change, the instructional psychologists have developed the concept of "corrective" feedback, defined as specifically directing the learner to modify an incorrect response (Anderson & Faust, 1973). Whereas knowledge of results assumes that probability of response will eventually change through conditioning, corrective feedback mandates changes.

The field of cybernetics has also adopted the feedback mechanism, especially through the modeling of human behaviors by computers or servo-mechanism simulation (Smith & Sussman, 1969). In the area of motor skills, feedback serves as a regulatory mechanism. Subsequently, the cybernetic feedback model has been extended to curriculum development (Pratt, 1981), and human-problem solving (e.g., Newell & Simon, 1976).

Cybernetics moves away from the incentive-with-information-backup model toward an information-as-primary position. The cognitive position also is that feedback has a far

greater impact on cognitive processing during the task, as compared to motivational consequences (Bourne, Guy, & Wadsworth, 1967; Spence, 1964; Estes, 1970).

The incentive-informational dispute is probably the most central issue, and has not been resolved. Indeed, it is likely that feedback combines both aspects, although the experimenter has the opportunity to suppress or enhance either dimension depending on design bias (Getsie, Note 5). As noted, in our paper we shall concentrate on the informational thrust.

Another major problem in neatly conceptualizing feedback is that it is basically a multi-dimensional concept (Holding, 1965). Using a branching paradigm, Holding postulated several dichotomies. The first creates a distinction between extrinsic and intrinsic feedback. Each of the first pair, in turn, can be subdivided into concurrent and terminal. Subsequent dichotomies include: immediate-delayed, verbal-non-verbal, and separate (after each response)-accumulated (after all responses). The number of possible permutations and combinations is awesome. A perusal of the dichotomies clearly indicates that feedback experiments are likely to persist in having the directive and incentive results confounded (Locke, Cantledge, & Kleppel, 1968), although as previously suggested by appropriate selection of treatment variables the incentive or information component can be suppressed.

Feedback, regardless of conceptual complexities and mixed research findings (e.g., Barringer & Gholson, 1979), is considered a critical component by most instructional systems

analysts and psychologists (e.g., Davis, Alexander, & Yelon, 1974). By any set of standards it is considered absolutely essential in designing auto-tutorial systems involving computer-based instruction (e.g., Gagne, Wager, & Rojas, 1981). The reason is that under conditions of individualized instruction the student is severely limted in terms of different sources of feedback, and hence the "equalization phenomenon" (Hilgard & Bower, 1975) is attenuated. That is, the student cannot make up for deficiencies anywhere in the instructional situation by utilzing a variety of informational sources. We are not arguing that teachers in traditional classrooms are better managers of feedback. Rather, in individualized technologically-assisted settings the student is obviously more restricted.

Behavioral psychologists, as well as those interested in the area of artificial intelligence (AI) have not been unmindful of this problem (Bunderson & Faust, 1976). Given the early failures in machine language, more extensive attempts were made to develop sophisticated programs which were capable of carrying out a tutorial dialogue with the student. One example was SOPHIE (Sophisticated Instructional Environment), built around problem-solving electronic circuitry (Brown, Burton, & Bell, Note 6).

The assumption was that by providing detailed and analytical feedback, errors could be markedly reduced, and the student redirected in his thinking (Brown et al., Note 5). While students may attempt to correct themselves with repeated opportunities (e.g., Singer & Pease, 1978), most instructional psychologists

point to findings of error perseveration (e.g. Kaess & Zeaman, 1960) as something to be avoided (e.g. Howe, 1970; Ladas, 1980). Ultimately, one supposes the goal is errorless learning (Terrace, 1963).

Unfortunately, recent work on tutorial languages suggests that the elaborated conceptual model may be detrimental to student performance (Gallagher, 1981). Using the BLOCKS program Gallagher provided diagnostic feedback and information based on an "expert" problem-solving model (which is typical of tutorial programs). However, students not given this help actually performed better on subsequent problems.

Sternberg and Ketron's (1982) previously cited work also suggests that a predetermined feedback strategy model may actually be providing the learner with irrelevant data not conducive to a functional, student generated strategy. It is ironic to note, when early assumptions about simplicity and directness of feedback in programmed instruction were invalidated by students short-circuiting the feedback process by copying answers (e.g. Kulhavy, 1977), subsequent attempts to overcompensate in the other direction may also have had deleterious effects. The key lies, of course, in the basic assumptions of a feedback model based on an "expert" or linear strategy versus those derived explicitly from an analysis of the individual learner's strategy. The difference is expressed also in terms of the curriculum model, with the curriculum developer on one hand deciding in advance what <u>should be</u> done as compared to a more functional analysis of what the individual learner <u>is</u>

doing. While the research evidence supporting either experimenter-or learner-determined sequences is mixed (Dansereau, Evans, Wright, Long, & Atkinson, Note 7), there is no reason for solely accepting the developer determined model.

There are several other methodological issues in feedback which we must consider in our research and findings. Foremost is the finding that disconfirmation (negative) feedback impacts achievement much more than confirmation (positive) feedback (Kulhavy, 1977). Researchers in concept acquisition (e.g. Bruner, Goodnow, & Austin, 1956; Trabasso & Bower, 1968) have found that learners shift strategy after disconfirmation, although not back to some simple random selection (Levine, 1966). The disconfirmation paradigm is critical to our study. More subtle and/or qualitative aspects of feedback have been generally ignored (Kulhavy, 1977), but will be part of this research.

The use of feedback during and after the acquisition phase presents several problems. If it is given too soon or is too readily accessible, the error rate increases (Sullivan, Schutz, & Baker, 1967). Moreover, delay in knowledge of results also has diverse effects. Kulhavy (1977) has argued that while feedback should be provided as often as possible during instruction, the content organization and difficulty of the material is also crucial. If the material is very difficult, students spend most of their time guessing at answers and trying to match feedback to the question (Kulhavy, Yekovich & Dyer, 1976). However, where the material is relatively easy and/or student confidence is high, students spend relatively little time on feedback. Battig

(1966), however, suggested that difficulty in learning might have postive consequences. Certainly the issues of when and how much feedback are significant.

Moreover, delay in feedback seems to sometimes facilitate acquisition (e.g., Surber & Anderson, 1975). The argument is that the delay allows for incorrect responses to be forgotten, which approximates the practice interference paradigm (Kulhavy & Anderson, 1972). Finally, very pertinent to our thereotical position and use of feedback is a study by Kulhavy & Parsons (1972) which suggests that the effect of feedback is minimal unless the instructional content is organized in some meaningful way to the learner (e.g., schema, script, etc.).

SECTION 3

This section deals with literature relevant to our text synthesis method. We should like to begin with a discussion of text analysis. It is not our intent to provide a detailed review of the text analysis method, but to point out the relevance of specific findings to our research model.

The text analysis approach has developed within a number of specialized subareas, including grammar, content, orienting tasks, and knowledge bundles. At the present moment, the theoretical trend is epistemological (i.e., how knowledge is structured). Current theories have been tied essentially to measures of recall (Reder, 1980).

We should emphasize that we are concerned with the acquisition process <u>in utero</u> and subsequent implications for education. Research has clearly demonstrated that recall for

prose is subject to the logic of the discourse (e.g., Frase, 1969), as well as the serial position of the discourse elements. There have been a number of attempts to categorize text processing (e.g., Dawes, 1964; Crothers, 1972; Frederiksen, 1975; Rumelhart, 1977; Meyer, 1975). Most have been characterized as more or less successful within the limits imposed by the specific model (Reder, 1980). Generally these programs attempted to simulate via computer the sequential reorganization of prose during processing.

More recently, attempts have been made to link acquisition to outcomes of schemata (or related constructs). These are usually defined as abstract representations of knowledge within the memory structure of the learner (e.g., Rumelhart & Ortony, 1977), and are considered an integral part of text processing. It is assumed that input can be facilitative or interfering depending on the interaction between schema and discourse logic. Relevant also to our research is Schank's (1972) study dealing with sentence and paragraph comprehension. Paragraph generation involves connecting sentences, which presupposes perhaps some integrative processing similar to Kintsch's (1974) micro-macrostructures. Nevertheless, it will be interesting to determine if there is a relative paucity of inferences available per sentence as Clark and Haviland (1976) suppose, and how the number of inferences increases with the reconstruction of discourse.

Subjective organization, as analyzed in text processing, is also a function of familiarity and meaningfulness. Indeed some

researchers (e.g., Anderson & Reder, 1979) argue that the schemata available to the subject may have more impact on acquisition than the logic of the discourse itself. The reader should keep in mind our previous discussion of subjective organization.

Comparatively little systematic research has been done using disconnected discourse. The result is that the findings lack coherence in terms of outcomes. Overall, using a free recall procedure, scrambled sequences generally yields poorer performance than connected discourse, regardless of unit size (Dansereau et al., Note 7).

In many experiments involving scrambled presentation little is done to assist the learner using disconnected materials. In a typical study (Frase, 1969), subjects using a name concept strategy were superior to those using an attribute strategy, and both were superior to the scrambled order group. Sasson (1971) interrelated two different topics, a variant of the scrambled order method, and found that a thematic organization produced superior student achievement as compared to temporal (i.e., sentences included a date) and logical arrangements.

Meaningfulness of the materials has also been investigated. Bruning (1970) embedded the test items in contexts which were relevant and ordered, relevant and scrambled, and irrelevant and scrambled. The findings are somewhat unusual in that there was no significant difference between the relevant and ordered and relevant and scrambled contexts, but both were superior to the irrelevant and scrambled. As in nearly all these studies the

discussion dealt with concept acquisitions in the ordered material; no one seemed to pay much attention to the what and how of achievement in the scrambled treatment. Tobias (1973), using programmed instruction, found that familiarity overcame to some extent the effects of scrambling. Again using programmed instruction, Brown (1970) found that students encountering a scrambled mathematics sequence took more time, made more errors, and performed more poorly on problems. This is to be expected in a subject area in which meaning is clearly tied to a hierarchical sequence.

The interrelatedness of the sentences is also a factor. Kissler and Lloyd (1973) found that highly interrelated sentences were more sensitive to scrambling. Rothkopf (1962) indirectly assessed this dimension when he used loosely organized sentences to create his discourse treatments. Anderson (1967) suggested that programmed instructionl units (i.e. the frame) are developed to be independent, and with the high redundancy of programmed instruction units scrambling may not be as destructive of performance.

While none of the studies cited a feedback mechanism, a study done by Schultz and DiVesta (1972) is unique in that students were allowed to take notes (an outside confirmation source). Again, while the emphasis was on the strategies employed with respect to organized research materials, the researchers also observed that students taking notes modified the passages to suit their needs. That is, they were not forced to at least initially accept an experimenter determined logic.

In summary, Bartlett (1932) demonstrated that strong recall tended to fit schemata. Current research (e.g. Kintsch, 1974; Meyer, 1975) has reaffirmed this proposition by demonstrating that the probability of inclusion within recall is related to the importance of the statement to the theme. Instructionally, the implications are significant, in that we have argued previously that organization is modified at time of input (Mandler, 1972). Hence, relevancy is a function of learner schemata. Behaviorally oriented instructional psychologists can develop alogorithms for organizing concept elements (e.g., Merrill & Tennyson, 1979) only by ignoring the concept of chunking (Miller, 1956; Buschke, 1976). Since it is assumed that feedback involves active memory (Anderson, 1980), both timing as well as congruence of feedback to schemata and strategy is crucial. These will be addressed in our research.

We feel it can be argued that in most educational situations, particularly in higher-order and non-hierarchical learning, feedback might better serve by assisting in self-correction (Anderson, 1980). This avoids the problem of the instructional developer intruding into functional, and perhaps unique, memory and processing strategies. Certainly we should take heed from Nisbett and Wilson's (1977) findings that learners often cannot accurately verbalize their own strategies. If the learner cannot do it, then instructional psychologists may indeed be on dangerous conceptual grounds. Hence, while Singer (1980) called for optimizing feedback, in the end it may be the learner that dictates the meaning of "optimal."

IV: RECOMMENDATIONS

It is rather difficult to specify in usual terms the research recommendations, since the project calls for an evolving research program analyzing several critical issues in feedback relevant to classroom instruction. Our major argument has been that the recipient of a prose discourse, whether written or spoken, builds and processes through a net of possible semantic constructions as the sentences are processed. Synthesis follows when sufficient linguistic cues have been received and transformed to a set of probable meanings, and the singular meaning of the text emerges. This processing is influenced partially by prior real-world knowledge which is idiosyncratic to the perceiver, as well as more commonly shared formal coding processes required by linguistic constraints (e.g. Bourne et al., 1979). Hence, indviduals can be expected to follow unique paths to the common semantic interpretation requred for linguistic competence. This hypothesized idiosyncratic processing is basic to the theoretical thrust of the proposed program of feedback research. To recapitulate, current feedback theory date involves both knowledge of results (i.e. confirming/disconfirming) and corrective feedback. It is our thesis that a minimal approach may be more effective, especially where learner processing is heavily involved.

The literature review demonstrates the many potential facets of our research program. However, the research plan will become better defined as our findings emerge. One of the crystallized paradigms will be discussed in detail in order to demonstrate

more explicitly the focus of the program. These experiments are necessary as an early exploration to demonstrate (1) the idiosyncratic nature of sentence processing, and (2) the possibility of manipulating feedback variables such as frequency and complexity to meet processing needs for the synthesis of the text schemata. The experiments involve the manipulation of feedback in relation to subjective organization, applying the Tulving paradigm to sentences rather than words.

Two stories based on simple active declarative sentences will be used so that verbatim recall as well as recognition may be clearly expected as outcomes. The stories will consist of 48 sentences, organized around the Battle of Gettysburg and the Superbowl. The discourse content logic will be established so there is high agreement among observers for a preferred order. Experimentally the stories can be presented separately or interrelated. The sentences will be presented at a six-second rate in random order at a CLIPR terminal, with the timing and sentence orders controlled by the computer.

The task of the learner will be to reconstruct the preferred order of the discourse. Each trial will be on a fixed-time basis, with a one-minute break between trials. There will be an upper limit to the number of trials per student. Both time per trial and number of trials will be derived from pilot studies. The feedback treatments will be discussed in greater detail as part of the experimental design.

EXPERIMENT 1

Design

After the first sentence is given, the reconstructed sentence sequence will remain continuously on screen, with the remaining sentences generated one at a time. Each sentence will actually have two numbers: One assigned by order within the story (unknown to the subject) and a presentation number given on the screen. These latter numbers will be sequential, in order of appearance to the subject. The subject will organize the discourse by reference to the screen numbers.

For example, let us suppose the first sentence that appears on the screen is in reality #10 in the story. On the screen the sentence appears as #1. Sentence #2 (in reality #3 in the story) then appears. If subjects assign Sentence #2 ahead of #1, they are correct. If they maintain #1 ahead of #2 they are in error. If Sentence #2 is placed ahead of Sentence #1, then #2 is now labeled "1", and #1 is labeled "2". Sentences will be automatically renumbered on the screen until the order matches the story.

Subjects will receive either story. The feedback content is defined in Experiment 1 as disconfirmati (i.e., the subject is informed that the placement of a sentence is in error). Subjects will also be told they may or may not receive feedback, to avoid the problem of assuming no feedback means "right." They will also be informed that they will be expected to recall the story. Feedback groups are assigned as follows:

Group I will receive no feedback on any trial.

Group II will receive feedback on the first error only within each trial.

Group III will receive feedback on all errors within each trial.

Group IV will receive feedback through the first 10 errors, regardless of trial.

Recall, recognition, time to criterion, and number of errors are the dependent variables. These will be described later.

Subjects will be randomly assigned to one of the two recall/recognition groups described below. The recall/recognition tests will be given 15 minutes after the last trial.

The resulting 2 (story) x 4 (feedback) completely crossed design permits an examination of the absolute versus relative frequency issue in feedback (Bourne et. al., 1967).

Dependent Variables

The dependent variables will include recall and recognition; 1) Recall: Subjects will be asked to recall in writing as much of the story as possible. Accuracy will be based on number of idea units.

2) Recognition (probe technique): Subjects will receive triads of sentences. Within each triad will be a sentence taken directly from the discourse, one a paraphrase, and one inappropriate (i.e. a foil). Accuracy scores will be based on number of verbatim discourse sentences chosen. Error analysis will also reflect number of paraphrase and irrelevant sentences.

Time to criterion and number of errors will be analyzed, and possibly covaried with the recognition and recall scores. This is true for all experiments.

EXPERIMENT 2

This experiment will manipulate the informational feedback and frequency issues. Again, each subject will receive one story, randomly assigned. The feedback groups will receive information as to process appropriate placement of a sentence when an error is made (as opposed to simply disconfirming placement).

Group I: Subjects receive no feedback.

Group II: Subjects receive informative feedback only on first error within each trial.

Group III: Subjects receive informative feedback on all error within trials.

Group IV: Subjects receive informative feedback on first 10 errors only, regardless of trial.

Again subjects will be assigned to either recognition or recall. The design is a 2 (story) x 4 (feedback) completely crossed model.

EXPERIMENT 3

This experiment will replicate portions of Experiment 1 (feedback as disconfirmation), but also assesses changes in feedback patterns on memory measures. Subjects will serially receive both stories counterbalanced across subjects. The memory measures (i.e., either recall or recognition) will be randomly

assigned after each story, counterbalanced across groups. Feedback (disconfirmation) groups will be assigned as follows:

Group I: Subjects receive no feedback on either story.

Group II: Subjects receive no feedback on one story, and first error per trial on the other.

Group III: Subjects receive feedback on first error only per trial for one story, and unlimited for other.

Group IV: Subjects receive no feedback on one story, and unlimited feedback on other.

The design is a 2 x 4 completely crossed model.

EXPERIMENT 4

This experiment will parallel the Experiment 3 paradigm, using informative feedback.

Additional Studies

Several other research issues are immediately apparent. Obviously, one area is to vary feedback within and between studies in terms of disconfirmation/information. A second issue is whether or not the experimenter tells subject when the order is complete, or allows the subject to terminate the process. Third, we would like to examine the results when the subject is asked to process the text on "gist" rather than recall instructions. Finally, we should begin to develop reliable instructional techniques for analzying strategies and incorporating them into the feedback model. It is hoped that these results will allow instructional developers to construct

feedback content and frequency which is more amenable to individual strategies. That is, it may be possible to "titrate" frequency and amount of information based on analyses of student strategies.

Reference Notes

1. Langer, P. An evaluation of selected course authoring systems. F33615-77-C-0038. Denver, Colorado: Air Force Human Resources Laboratory, Lowry Air Force Base, 1978.

2. Posner, G.J. & Strike, K.A. An analysis of curriculum structure. Paper presented at the meeting of the American Educational Research Association, Chicago, 1974.

3. Kozminsky, E., Kintsch, W., Coreu, P., and Bourne, L.E. Comprehension and analyses of information in text: II. Decision making with texts. (Technical Report No. 89-ONR), Boulder, Colorado: Institute for the Study of Intellectual Behavior, 1979.

4. Clark, R.E. Instructional methods: Operational definitions and critical attributes. Paper presented at the meeting of the American Educational Research Association, New York, March, 1982.

5. Getsie, R. Meta-analysis of the effects of type and combination of feedback on children's discrimination learning. (Doctoral dissertation, University of Colorado, 1982).

6. Brown, J.S., Burton, L.R., & Bell, A.A. SOPHIE: A sophisticated instructional environment for teaching electronic troubleshooting (An example of AI in CAI). F-41609-73-C-0006. Lowry Air Force Base, Denver, Colorado: Air Force Human Resources Laboratory, 1974.

7. Dansereau, D.R., Evans, S.H., Wright, A.D., Long, G., & Actkinson, T. Factors related to developing instructional information sequences: Phase I. AFHRL TR73-51(1). Brooks Air Force Base, San Antonio, Texas: Air Force Human Resources Laboratory, 1973.

References

Anderson, J.R. Cognitive psychology and its implications. San Francisco: W.H. Freeman & Co., 1980.

Anderson, J.R., & Reder, L.M. An elaborative processing explanation of depth of processing. In L.S. Cermak & F.I.S. Craik (Eds.), Levels of processing human memory. Hillsdale, New Jersey: Lawrence Erlbaum, 1979.

Anderson, R.C., & Faust, G.W. Educational psychology: The science of instruction and learning. New York: Dodd, Mead & Co., 1973.

Anderson, O.R. The strength and order of responses in a sequence as related to the degree of structure in stimuli. Journal of

Research in Science Teaching, 1966, 4, 192-198.

Atkinson, R.C., & Wickens, T.D. Human memory and the concept of reinforcement. In R. Glaser (Ed.), The nature of reinforcement. New York: Academic Press, 1971.

Barringer, C., & Gholson, B. Effects of type and combinations of feedback upon conceptual learning by children: Implications for research in academic learning. Review of Educational Research, 1979, 49, 459-478.

Bartlett, F.C. Remembering. Cambridge, England: Cambridge University Press, 1932.

Battig, W.F. Facilitation and interference. In E.A. Bilodeau (Ed.), Acquisition of skill. New York: Academic Press, 1966.

Bilodeau, E.A. Supplementary feedback and instructions. In E.A. Bilodeau & I.McD. Bilodeau (Eds.) Principles of skill acquisition. New York: Academic Press, 1969.

Bloom, B.S. Learning for mastery. In B.S. Bloom (Ed.), All our children learning: A primer for parents, teachers, and other educators. New York: McGraw-Hill, 1981.

Bourne, L.E., Jr., & Restle, F. A mathematical theory of concept identification. Psychological Review, 1959, 66, 278-296.

Bourne, L.E., Jr., Guy, D.E., & Wadsworth, N. Verbal-reinforcement combination in a card-sorting task. Journal of Experimental Psychology, 1967, 73, 220-224.

Bourne, L.E., Jr. Knowing and using concepts. Psychological Review, 1970, 77, 546-556.

Bourne, L.E., Jr., Dominoswki, R.L., & Loftus, E. Cognitive processes. Englewood Cliffs, New Jersey: Prentice-Hall, 1979.

Bousfield, W.A. Occurrence of clustering in recall of randomly arranged units. Journal of General Psychology, 1953, 49, 229-240.

Brown, J.L. Effects of logical and scrambled sequences in mathematical materials in learning with programmed instruction. Journal of Educational Psychology, 1970, 61, 41-45.

Bruner, J.S., Goodnow, J., & Austin, G. A study of thinking. New York: John Wiley, 1956.

Bruning, R.H. Short-term retention of specific factual information in prose contexts of varying organization and relevance.

Journal of Educational Psychology 1970, 61, 186-192.

Bunderson, C.V., & Faust, G.W. Programmed and computer-assisted instruction. In N. Gage (Ed.), The psychology of teaching methods: The Seventy-fifth Yearbook of the National Society for the Study of Education (Part I). Chicago: University of Chicago Press, 1976.

Buschke, H. Learning is organized by chunking. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 313-324.

Buss, A.H., Braden, W., Orgel, A., & Buss, E. Acquisition and extinction with different verbal reinforcement combinations. Journal of Experimental Psychology, 1956, 52, 288-295.

Clark, H.H., & Haviland, S.E. Comprehension and the given-new contract. In R.O. Freedle (Ed.), Discourse production and comprehension. Hillsdale, New Jersey: Lawrence Erlbaum, 1976.

Crothers, E. Memory structure and the recall of discourse. In R. Freedle & J.B. Carroll (Eds.). Language comprehension and the acquisition of knowledge. Washington, D.C.: Winston, 1972.

Davis, R.H., Alexander, L.T., & Yelon, S.L. Learning system design: An approach to the improvement of instruction. New York: McGraw-Hill, 1974.

Dawes, R. Cognitive distortion. Psychological Reports, 1964, 14, 443-459.

Earnhard, M. Free recall, transfer, and individual differences in subjective organization. Journal of Experimental Psychology, 1974, 103, 1169-1174.

Estes, K. An information-processing analysis of reinforcement in children's discrimination learning. Child Development, 1976, 47, 639-647.

Frase, L.T. Paragraph organization of written materials: The influence of conceptual clustering upon level of organization. Journal of Educational Psychology, 1969, 60, 394-401.

Fredericksen, C.H. Representing logical and semantic structure of knowledge acquired from discourse. Cognitive Psychology, 1975, 7, 139-166.

Gage, N.L., & Berliner, D.C. Educational Psychology (2nd ed.). Chicago: Rand McNally, 1979.

Gagne, R.M. The conditions of learning (3rd ed.). New York: Holt, Rinehart, and Winston, 1977.

Gagne, R.M., Wager, W., & Rojas, A. Planning and authoring computer-assisted instruction lessons. Educational Technology, 1981, 21, 17-26.

Gallagher, J.P. The effectiveness of man-machine tutorial dialogue for teaching attribute blocks problem-solving skills with an artificial intelligence CAI system. Instructional Science, 1981, 10, 297-332.

Glidden, L., & Roemer, J. Meaningfulness, formal similarity, and subjective organization. Psychological Review, 1974, 24, 399-408.

Gosfern, D., & Blair, C. Factors effecting multitrial free recall. Journal of Educational Psychology, 1971, 62, 17-24.

Haviland, S.E., & Clark, H.H. What's new? Acquiring new information in the process of comprehension. Journal of Verbal Learning and Verbal Behavior, 1975, 13, 512-521.

Hilgard, E.R., & Bomer, G.H. Theories of learning (4th ed.). New York: Appleton-Century Crofts, 1975.

Howe, M.J.A. Using students' notes to examine the role of the individual learner in acquiring meaningful subjective matter. Journal of Educational Research, 1970, 64, 61-63.

Kaess, W., & Zeaman, D. Positive and negative knowledge of results on a Pressey-type punchboard. Journal of Experimental Psychology, 1960, 60, 12-17.

Keller, F.S. "Good-bye, teacher". Journal of Applied Behavioral Analysis, 1968, 1, 78-89.

Kintsch, W. The representation of meaning in memory. Hillsdale, New Jersey: Lawrence Erlbaum, 1974.

Kissler, G.R., & Lloyd, K.E. Effect of sentence interrelation and scrambling on the recall of factual information. Journal of Educational Psychology, 1973, 64, 187-190.

Kulhavy, R.W., & Parsons, J.A. Learning-criterion error presentation in text materials. Journal of Educational Psychology, 1972, 63, 81-86.

Kulhavy, R.W., & Anderson, R.C. Delay-retention effects with multiple-choice tests. Journal of Educational Psychology, 1972, 63, 505-512.

Kulhavy, R.W., Yekovich, F.R., & Dyer, J.W. Feedback and response confidence. Journal of Educational Psychology, 1976, 68, 522-528.

Kulhavy, R.W. Feedback in written instruction. Review of Educational Research, 1977, 47, 211-232.

Ladas, H. Summarizing research: A case study. Review of Educational Research, 1980, 50, 597-624.

Levine, M. Hypothesis behavior by humans during discrimination learning. Journal of Experimental Psychology, 1966, 71, 331-338.

Levine, M. A cognitive theory of learning. Hillsdale, New Jersey: Lawrence Erlbaum, 1975.

Locke, E.A., Cartledge, N., & Koeppel, J. Motivational effects of knowledge of results: A goal-setting phenomenon? Psychological Bulletin, 1968, 70, 474-485.

Maltzman, I. Thinking: From a behavioristic point of view. Psychological Review, 1955, 66, 367-386.

Mandler, G. Organization and recognition. In E. Tulving & W. Donaldson (Eds.), Organization of memory. New York: Academic Press, 1972.

Markle, S.M. They teach concepts, don't they? Educational Researcher 1975, 4, 3-9.

Merrill, D.M., & Tennyson, R.D. Teaching concepts: An instructional design guide. Englewood Cliffs, New Jersey: Educational Technology Program Publications, 1977.

Miller, G.A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 1956, 63, 81-97.

Myer, B.J.F. The organization of prose and its recall. Amsterdam: North-Holland Publishers, 1975.

Newell, A., & Simon, H. Human problem solving. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

Nisbett, R.E., & Wilson, T.D. Telling more than we can know: Verbal reports on mental processes. Psychological Review, 1977, 84, 231-259.

Pellegrino, J. A Fortran IV program for analyzing higher order subjective limits in free recall learning. Behavioral Research Methods and Instruction, 1972, 4, 215-217.

Pellegrino, J., & Battig, W.F. Relationship among higher order organizational measures and free recall. Journal of Experimental Psychology, 1974, 102, 463-472.

Postman, L. A pragmatic view of organization theory. In E.

Tulving & W. Donaldson (Eds.), Organization of memory. New York: Academic Press, 1972.

Pratt, D. A cybernetic model for curriculum development. Instructional Science, 1981, 11, 1-12.

Reder, L.M. The role of elaboration in the comprehension and retention of prose: A critical review. Review of Educational Research, 1980, 50, 5-53.

Rumelhart, D.E. Understanding and summarizing brief stories. In D. LaBerge & J. Samuels (Eds.), Basic processes in reading: Perception and comprehension. Hillsdale, New Jersey: Lawrence Erlbaum, 1977.

Rumelhart, D.E., & Ortony, A. The representation of knowledge in memory. In R.C. Anderson, R.J. Spiro, & W.E. Montague (Eds.)., Schooling and the acquisition of knowledge. Hillsdale, New Jersey: Lawrence Erlbaum,, 1977.

Robin, A.L. Behavioral instruction in the classroom. Review of Educational Research, 1976, 6, 313-354.

Rosch, E.H., & Mervis, C.B. Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 1975, 7, 573-605.

Rothkopf, E.Z. Learning from written sentences: Effects of order of presentation on retention. Psychological Reports, 1962, 10, 667-674.

Sasson, R.Y. Semantic organizations and memory for related sentences. American Journal of Psychology, 1971, 84, 253-267.

Schultz, C.B., & DiVesta, F.J. Effects of passage organization and note taking on the selection of clustering strategies on recall of textual materials. Journal of Educational Psychology, 1972, 63, 244-252.

Shank, R.C. Conceptual dependency: A theory of natural language and understanding. Cognitive Psychology, 1972, 3, 552-631.

Shuell, T. Clustering and organization in free recall. Psychological Bulletin, 1969, 72, 352-374.

Singer, R.N., & Pease, D. Effects of guided versus discovery learning strategies on initial motor task learning, transfer, and retention. American Alliance for Health, Physical Education and Recreation Research Quarterly, 1978, 49, 206-217.

Singer, R.S. Motor learning and human performance: An application to motor skills and movement behavior. New York: MacMillan, 1980.

Skinner, B.F. Science and human behavior. New York: MacMillan, 1953.

Smith, K.V., & Sussman, H. Cybernetic theory and analysis of motor learning. In E. Bilodeau & I.McD. Bilodeau (Eds.), Skill acquisition. New York: Academic Press, 1969.

Spence, J.T. Verbal discrimination performance under different verbal reinforcement combinations. Journal of Experimental Psychology, 1964, 67, 195-197.

Sullivan, H.J., Schutz, R.E., & Baker, R.L. Effects of intrinsic and extrinsic reinforcement contingencies on learner performance. Journal of Educational Psychology, 1967, 58, 165-169.

Surber, J.R., & Anderson, R.C. Delay-retention effects in natural classroom settings. Journal of Educational Psychology, 1975, 67, 170-173.

Terrace, H.S. Discrimination learning with and without "errors". Journal of the Experimental Analysis of Behavior, 1963, 6, 1-27.

Thorndike, E.L. The psychology of learning: Education psychology (Vol. II). New York: Teachers College, Columbia University, 1913.

Thorndike, E.L. The fundamentals of learning. New York: Teachers College, Columbia University, 1932.

Tobias, S. Sequence, familiarity, and attribute by treatment interactions in programmed instruction. Journal of Educational Psychology, 1973, 64, 133-141.

Trabasso, T., & Bower, G.H. Attention in learning. New York: John Wiley, 1968.

Tulving, E. Subjective organization in the free recall of unrelated words. Psychological Review, 1962, 69, 344-354.

Winne, P.H. Minimizing the black box problem to enhance the validity of theories about instructional effects. Instructional Science, 1982, 11, 13-28.

Wood, G. Organizational processes and free recall. In E. Tulving & W. Donaldson (Eds.), Organization of memory. New York: Academic Press, 1972.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

THE DETECTION OF HAZARDOUS MATERIALS FROM SPILLS

Prepared by: Dr Stephen F. Lin

Academic Rank: Associate Professor

Department and University: Department of Chemistry
North Carolina Central University

Research Location: Air Force Rocket Propulsion Laboratory
Liquid Rocket Division
Liquid Propellant Systems Section

USAF Research Colleague: Mr Forrest S. Forbes

Date: July 30, 1982

Contract No.: F49620-82-C-0035

THE DETECTION OF HAZARDOUS MATERIALS FROM SPILLS

by

Stephen F. Lin

ABSTRACT

This report discusses the instrumentation for detection and identification of hazardous materials. Both remote sensing instruments and non-remote portable instruments have potential application in emergency operations. The remote sensing techniques are reviewed as to their principles and their applications in air pollution measurements. Commercially available semi-portable, portable, and personal instruments are surveyed, and the principles of operation are briefly discussed. The instruments that are being developed for hypergol vapor detection are specifically used.

I. INTRODUCTION:

The Air Force Rocket Propulsion Laboratory, under contract to Systems Technology Laboratory, Inc., is currently performing a state-of-the-art assessment of technology which will minimize hazards and environmental damage from transportation-related accidents or other spills of certain chemicals and propellants. The ultimate goal of this project is to develop procedures and guidelines for emergency response, hazards mitigation, wreckage removal, and cleanup and disposal of hazardous material spills. One of the most important problems in the post accidental procedures is the identification and detection of the hazardous materials involved. Due to the time limit of the Summer Faculty Research Program, and the special interest of the Rocket Propulsion Laboratory, I decided to concentrate my research only in the instrumentation for detecting hazardous vapors, especially on the instruments that can be used in the emergency operations.

The ideal instrument for identifying and monitoring the hazardous vapors from spills would be a lightweight portable instrument, which could be brought near, but at a distance from the scene, and be operated by any inexperienced persons, and which could be pointed at the vapor cloud thus immediately identifying the vapor and determining the concentration. Unfortunately, no such instrument has ever been developed. The ideal instrument mentioned above would be a combination of a remote sensing instrument and a portable gas detector. Therefore, I studied various remote sensing techniques as applied to air pollution monitoring, and gathered information about commercially available gas detectors or analyzers. As the remote sensing technology progresses, an instrument may be developed that can detect hazardous vapor rapidly, accurately, and remotely.

II. OBJECTIVES:

The objectives of my study were as follows:

1. To study the various remote sensing techniques currently being developed for pollution monitoring, atmospheric studies, space exploration, and chemical warfare.

2. To survey detection devices and portable instruments that are on the market for possible use in detection and monitoring of hazardous chemicals and propellants.

III. REMOTE SENSING TECHNOLOGY:

Remote sensing can be defined as sensing qualitatively and/or quantitatively a species in the environment where the instrument and the species under investigation do not come into direct contact. This would be a desirable characteristic for the identification and monitoring of hazardous materials from spills. Unfortunately, most of the remote sensing techniques being developed use powerful lasers as light sources, require sophisticated electronics and computers for data collection and analysis, need skilled personnel to operate the instrument and interpret the data, and usually can measure only one or a few pre-determined species. These features would limit the applicability of the remote sensing technique to the identification and monitoring of hazardous emitted materials from spills at the present time. Nevertheless, the development in the laser remote sensing technology may someday lead to the application of these techniques in the remote monitoring of hazardous materials.

Remote sensing techniques utilize different interacting mechanisms between the radiation and the species under investigation to infer the identity and the concentration of the species. They can be divided into two broad categories: active and passive, depending on the source of radiation. If the light source is a non-natural light source such as a laser, it is an active system. If the light source is a natural light source such as sunlight, it is a passive system. Active systems can be double-ended or single-ended depending on whether the light source and the detector are separated or at the same location. Active techniques include direct absorption, photoacoustic techniue, light detection and ranging (LIDAR), differential absorption and scattering (DAS or DIAL), Raman scattering (including resonance Raman scattering), laser induced fluorescence, and heterodyne detection. Passive techniques include passive infrared absorption and emission, and correlation technique. Each of these techniques will be discussed below as to their principles and applications in the air pollution measurements. A number of books or review articles have been written on the general subject.[2-5]

1. Direct Absorption

Direct absorption is the simplest and most sensitive remote monitoring technique. In this technique, radiation passing through the target species is absorbed by the species and detected by a detector. The absorption spectrum shows a characteristic pattern of the absorbing species. The radiation may be a continuous infrared, visible, or ultraviolet light source, or a multi-wavelength tunable laser. The radiation may be passed through an interferometer before being detected, giving an interferogram which can be made to yield a conventional absorption spectrum through Fourier transformation. This is called the Fourier transform interferometer system which gives higher resolution than the conventional absorption system.

There have been many applications of the direct absorption technique in the remote detection of air pollutants using non-laser light source,[5] Fourier transform spectrometer system,[6,7] or tunable lasers.[8,9]

2. Photoacoustic Technique

The photoacoustic (or opto-acoustic) technique measures the change in the pressure of the sample when the sample absorbs the radiation. The pressure fluctuation can be changed to acoustic signal by a sensitive microphone. The acoustic signal is proportional to the intensity of the radiation and to the concentration of the absorbing species.[10]

The use of opto-acoustic technique for air pollutant detection was proposed only recently by Kreuzer", but it has become a very useful technique for the detection of many air pollutants.[12-14]

3. Light Detection and Ranging (LIDAR)

As the light (usually a laser) passes through the atmosphere, the gas molecules and particulates cause elastic scattering (Rayleigh and Mie scattering) of the radiation. A small fraction of the light is backscattered in the direction of the LIDAR system and can be collected by a telescope and detected by a sensitive photomultiplier. From the signal received as a function of time and its intensity, the presence, range, and concentration of particulate clouds can be determined.

LIDAR is the most effective method for remote detection of smoke plume and dust clouds. Many LIDAR systems, both ground-based or airborne have been built and tested, and their success in the determination of cloud distribution has been demonstrated.[15-17]

4. Differential Absorption and Scattering (DAS) or Differential Absorption LIDAR (DIAL)

The technique utilizes two pulsed lasers which are tuned to two specific wavelengths corresponding to the maximum and minimum in the absorption spectrum of the particular gas of interest. As the laser travels through the atmosphere, Rayleigh and Mie backscattered radiation is measured as a function or range. Difference in the backscattered laser intensity from the two transmitted wavelengths is due to the difference in absorption by the gas. Analysis of the detected signals from both beams as a function of time provides a range-resolved measurement of the concentration of the gas.

Since Schotland measured the water vapor vertical profiles in his pioneer work,[18] this technique has been applied to measure very low concentration of pollutant gases over long distances.[19,20] An airborne unit can provide a three-dimensional map of pollutant over an area from the aircraft to the ground.[21]

5. Raman Scattering (Including Resonance Raman Scattering)

When a monochromatic light passes through a sample, it is inelastically scattered by the molecules. Its frequency is shifted slightly higher or lower due to the change in the vibrational and rotational energy states of the molecules. This is called the Raman scattering, and the spectrum is called the Raman spectrum. The magnitude of the shift is unique to the scattering molecule and the intensity of the Raman band is proportional to the concentration of the sample. This is the basis of the Raman scattering technique. The Raman (inelastic) scattering cross section is about three orders of magnitude smaller than the corresponding Rayleigh (elastic) scattering cross section. Therefore, its application in remote sensing was not possible until very powerful lasers became available. When the frequency of the incident radiation approaches the frequency of an allowed transition in a molecule, the Raman scattering cross section increases as much as three orders of magnitude over the ordinary Raman scattering.

This is resonance Raman scattering. The technique based on the resonance Raman scattering should increase the sensitivity, but it became possible only recently when the tunable laser sources became available.

Inaba et al. have done many remote analyses of molecular pollutants since 1969 using the Raman scattering scheme.[22-24] Others have also used this scheme for remote analysis of atmospheric H_20, $C0_2$, SO_2, and kerosene vapor.[25-26] The technique based on the resonance Raman scattering scheme has also been applied to the measurements of I_2, NO, NO_2, Cl_2 and O_3.[27]

6. Laser-Induced Fluorescence

Fluorescence is the spontaneous emission of a photon after an atom or a molecule has been excited to an excited state by absorption of incident radiation at a frequency within a specific absorption line of the species. The re-emitted radiation is useful in identifying and monitoring the species responsible for the fluorescence. For this technique, tunable dye lasers are probably most suitable.

There have been many discussions of the application of this technique in remote detection, but most of the studies have been done in the laboratory. Both atomic species like Na, and Hg,[28,29] and molecular species like CO, SO_2, NO, NO_2, I_2,[28,29] have been studied by this technique.

7. Heterodyne Detection

The heterodyne detection technique is the most sensitive method for detecting weak radiation signals. In this technique, a laser is used as a local oscillator, and its signals mixed with the signal from the sample of interest, so that the difference in the frequencies can be detected by a non-linear detector. These signals can be thermal radiation from gases or laser radiation scattered from gas molecules or aerosols. Therefore, the heterodyne detection technique has been used in passive remote detection using sun as a blackbody source to measure atmospheric constituents like ozone.[34] It has been applied in conjunction with infrared absorption,[35-36] differential absorption,[37] and laser-induced fluorescence[38] to increase the detection sensitivity.

8. Passive Infrared Spectroscopy

Active infrared spectroscopy has long been proved as the most useful tool for gas analysis in the laboratory. Infrared spectroscopy in passive mode has only been used recently for gas analysis since the commercial Fourier transform spectrometers became available.[39] The most important feature of the passive infrared spectroscopy is its potential for remote detection of gases. In order for this technique to be applicable, it is necessary that the background and the gaseous body be at different temperatures. When the background is at a higher temperature than the gas, the blackbody radiation from the background is at a higher temperature than the gas; the blackbody radiation from the background is absorbed by the gas and is observed. When the gas is at a higher temperature, the emission of the gas is observed.

Both the absorption and emission modes have been used for remote gas analysis. The U.S. Army has been developing a remote sensing system called the XM21 for the detection of nerve gases.[40] It is operated in the absorption mode and is designed to detect and identify a certain specific gas from a small number of substances using filtering and correlation techniques. NASA has developed a number of remote detection systems for space vehicles based on the passive infrared absorption technique for atmospheric studies using the sun as the source. The passive infrared emission spectroscopy has been applied to study waste gas flares and hot gas plumes from jet engines or power plants.[6,39,41]

9. Correlation Technique

The correlation technique is often used with UV absorption or IR emission spectroscopy to detect a specific gas from several species. In the UV absorption, light from sky passing through the gas cloud is collected by the instrument. It is then dispersed into its spectral components which are transmitted selectively through filters or correlation devices. In the IR emission, the emission radiation is viewed through a series of windows each fitted with a filter to establish a spectral bandpass centered on an emission line of a gas of interest. This correlation technique has been applied for air pollutant detection.[42-44] It offers the most promising remote sensing technique for hazardous materials.

IV. COMMERCIAL GAS DETECTORS AND ANALYZERS

Although remote sensing is very ideal and desirable for detecting and monitoring hazardous vapors from spills, there are some serious problems involved in the application of remote sensing techniques in hazardous material detecting as mentioned in the last section. Those problems have to be solved before the techniques can be used routinely for hazardous material spill detection. Except for a few mobile units and the Army's XM21 remote sensing system, it probably would take a while for most of the remote sensing techniques to be fully developed. Therefore, at the present time, the commercial instruments may serve the purpose of detecting and identifying hazardous materials from spills. Most of the commercial instruments are designed for air quality monitoring, gaseous pollutant detection, and fire or other emergency applications. Some of them are quite sophisticated and can make very accurate measurements and yet are lightweight and may be easily transported from location to location. Many of them are portable in the form of a kit, or even the personal type in the form of a badge or dosimeter. Most of the portable instruments or personal devices are designed to detect one or a few specific pollutant gases. Also, many of them can be modified to detect hydrazine fuels or other hypergol vapors. In fact, a few manufacturers have been developing detectors for hypergol vapors,[45] and these detectors are being tested and evaluated by the Air Force and NASA.[46]

In the following paragraphs, first, the principal methods by which the instruments operate will be briefly reviewed. Then, the commercial instruments will be surveyed according to three groups: (1) the bench or rack mounted laboratory instruments that can be moved from one location to another, (2) the truly portable type instruments that are lightweight, operated by battery, and can be carried by a person, (3) personal devices that are wearable such as dosimeters or badges. The review of the methods of operation and the survey of the commercial instruments are not intended to be exhaustive.

1. Principal Methods of Operation

A. Absorption Spectroscopy

When an electromagnetic radiation is passed through the sample, the photons of certain frequencies are absorbed by the molecules of the sample.

The wavelengths of absorption are characteristic of the sample and the intensity of absorption, at a certain characteristic frequency, is a measurement of the concentration of the sample. This is the principle by which the absorption spectroscopy is used to identify and measure the concentration of an unknown sample. The electromagnetic radiation can be infrared, visible, or ultraviolet radiation, but infrared is the simplest and the cheapest. The IR gas analyzers manufactured by Infrared Industries, Inc., Model 8641865 IR analyzers by Beckman Instruments, Inc., the MIRAN Series gas analyzers of the Foxboro Company, and the Photomet 3100 UV photometric ozone analyzer of Columbia Scientific Industries Corporation are based on this principle.

B. Mass Spectrometry (MS)

When the molecules of the gas are bombarded by high energy electrons, they are ionized. The ions are accelerated by an electric potential and then passed through a magnetic field. They are separated by the speed and the mass-to-charge ratio of the ions. This is the general principle of the mass spectrometry. One type of mass spectrometers is the time-of-flight (TOF) mass spectrometer in which the accelerated ions are allowed to pass through a drift region. The distance that an ion travels depends on the mass of the ion. The CVE Superspec 600 gas analyzer operates on this principle. Another type of mass spectrometers is the quadrupole mass spectrometer. The ions are introduced into the quadrupole mass filter where an electric field is established between four parallel rods. The trajectory of an ion through the field depends on quadrupole electric field parameters and on the mass-to-charge ratio of the ion. By selecting appropriate field parameters, specific ions can be passed through the filter and be detected. The CVE QUAD quadrupole mass analyzer operates on this principle.

C. Gas Chromatography

Gas Chromatography is a separation method whereby the components of a mixture are separated from one another by the selective process of distribution between two immiscible phases; one is the mobile phase and the other is the stationary phase. If the mobile phase is a gas (usually referred to as carrier gas), it is called the gas chromatography. The gas chromatography can be either gas-liquid or gas-solid chromatography depending on whether the stationary phase is a liquid or a solid. The separated gas phase bands are detected by a detection device. A variety of detectors can be used, such as mass spectrometers,

flame ionization detectors, thermal conductivity detectors, and gas density detectors. The GOW-MAC Instrument Company manufactures a number of gas chromatographs using different detection systems.

D. Flame Ionization Detection (FID)

In the flame ionization detection, the sample is mixed with hydrogen and burned in air or oxygen. As ions and electrons are formed in the flame, they enter the gap between two electrodes across which a potential exists. This lowers the resistance across the gap and causes a current to flow. The current is directly proportional to the concentration of the sample gas in the flame. A number of instruments use the flame ionization detection principle. They include the GOV-MAC 750 Series GC, Columbia Scientific Industries Corporation's hydrocarbon analyzers, Beckman Instruments' hydrocarbon analyzers.

E. Thermal Conductivity Detection (TCD)

Thermal conductivity detection system uses the Wheatstone Bridge principle. Two pair of heating elements, placed in cavities within a metal block thermal conductivity cell, act as two legs at the bridge. A constant voltage DC power supply furnishes current to heat the filaments. One pair of the cell cavities contain a reference gas of known thermal conductivity. The remaining two cavities contain the sample gas. Difference in the thermal conductivity of the reference and sample gases causes an unbalance in the temperature and, therefore current requirement of the bridge heating filaments. The degree of unbalance is directly proportional to the quantity of the sample gas. Some of the GOW-MAC gas chromatograph and the GOW-MAC portable gas analyzers and gas leak detectors use the thermal conductivity detection principle. Beckman Instruments' Model 7-C also uses the same principle for gas analysis.

F. Flame Photometric Detection (FPD)

When the sample is burned in a hydrogen rich air flame, the atom acquires the thermal energy from the flame and is excited to a level at which it will radiate its characteristic emission spectrum. The intensity and frequency of the radiaiton are measured by photo-multiplier. This allows qualitative and

quantitative analysis of the atom present in the sample. This is the principle of the flame photometric detection on which the Columbia Scientific Industries Corporation's total sulfur, SO_2, H_2S, phosphorous anlyzers operate.

G. Fluorescence

Fluorescence is a type of luminescence in which an atom or a molecule emits visible radiation in passing from a higher to a lower electronic state. The intensity of radiation depends on the concentration of the atom or molecule. Meloy Laboratories and Beckman Instruments' Fluorescrnce SO_2 analyzers are based on this phenomenon.

H. Chemiluminescence

If the electronic excitation process in the luminescence is due to a chemical reaction, it is called chemiluminescence. A number of detectors are based on this phenomenon. For example, the nitric oxide detectors are based on the chemiluminescent reaction between NO and O_3 which produces photons in the 600 to 2500 nm region. The ozone detectors are based on the chemiluminescent reaction between ozone and ethylene molecules which produce light in the 300 to 600 nm region. The hydrazine chemiluminescence analyzer being developed by Aero Chem Laboratories is also based on the chemiluminescent reaction between hydrazines and ozone.

I. Electrochemical Methods

Some gas analyzers are based on electrochemical reactions that take place in an electrolytic cell. For example, the Interscan Corporation's gas analyzers utilize a gas-diffusion type electro-catalytic sensing electrode and a combination reference-counter electrode immersed in an electrolyte chamber. The gas molecules from the sample are absorbed on the sensing electrode where they undergo an oxidation or reaction by the proper selection of an applied constant potential difference between the sensing and reference electrodes. The reaction generates an electric current directly proportional to the gas concentration.

J. Change in Resistance Method

There are a large number of combustible gas indicators on the market. They all operate on the change in resistance method. A filament is connected as an arm of the balanced Wheatstone Bridge and operated at controlled temperature. The combustible gas burns in the filament, raises its temperature, and changes its resistance, unbalancing the bridge. The amount of unbalance indicates the concentration of the conbustible gas.

K. Colorimetric Method

When the sample reacts with certain proper chemical reagents, a color may result. The intensity of the color or the length of color change in a detector tube or paper tape indicates the concentration of the sample. The MSA detector tubes and the Du Pon's colormetric air monitoring badges are based on this method.

2. Commercial Instruments

A. Bench or rack-mounted laboratory instruments

Infrared Industries, Inc. [47]

Several models of non-dispersive infrared analyzers can detect a single gas or two gases.

GOW-MAC Instrument Co.[48]

Several models of gas chromatograph using flame ionization detection, thermal conductivity detection, or gas density detection system

Meloy Laboratories, Inc./Columbia Scientific Industries, Corp.[49]

SO_2, H_2S, total sulfur, hydrocarbon analyzers; $NO/NO_2/NO_x$ detectors; O_3 detector and analyzer; phosphorous area monitor

The Foxboro Company [50]

MIRAN Series infrared spectrometers for gas or vapor analysis; the company also developed a specialized hydrazine-fuel/NH_3 detector under Naval Research Laboratory technical guidance on behalf of NASA[45]

CVC Products, Inc. [51]

CVC QUAD Quadrupole mass analyzer and Superspec 600 TOF mass spectrometer

Beckman Instruments, Inc. [52]

Non-dispersive IR gas analyzers; thermal conductivity gas analyzers; O_2, CO, SO_2, O_3, NO/NO_2, hydrocarbon analyzers.

Aero Chem Research Laboratories, Inc. [53]

Under contract with the School of Aerospace Medicine at Brooks AFB, TX, is developing a chemiluminescence detector for the ambient air detection of hydrazines and NO_2 [45]

The Aerospace Corporation [54]

Developing a CO_2 laser photoacoustic detector for hydrazine fuels and other toxic vapors[45]

B. Portable Gas Detectors

Mine Safety Appliance (MSA) Company [55]

Chlorine, combustible gas, hydrocarbon, oxygen, CO indicators and alarms; Universal Tester; detector tubes

Interscan Corporation [56]

Portable gas analyzers for CO, H_2S, SO_2, NO_2, NO, Cl_2; is currently under contract to the school of Aerospace Medicine at Brooks AFB, TX to develop a battery-operated, hand-held voltammetric instrument for ambient air detection of hydrazines [45]

GOW-MAC Instrument Co [48]

Gas leak detectors, portable gas analyzers, CO_2 monitor

Infrared Industries, Inc.[47]

O_2, flue gas, CO_2, CO, hydrocarbon analyzers

The Mogul Corporation

Portable gas analyzers for SO_2, NO_x, CO, O_2, total sulfur

The Foxboro Company [50]

MIRAN-101, 103, 104 portable vapor or gas analyzers

Hach Company [58]

Hach Company manufactures a hazardous material detection kit that contains chemical reagents, PH meter, conductivity meter, indicators, miniature chromatographic column, spectrophotometer, etc., and can do various tests for toxic substances.

Bacharach Instruments [59]

Mercury vapor, oxygen, combustible gas, natural gas, H_2S, CO monitors or sniffers.

Beckman Instruments, Inc.[52]

Portable oxygen, CO analyzers

Energetics Sciences, Inc.[60]

ESI is under contract to NASA to develop instrumentation to detect the hydrazine fuels and NO_2.[45]

International Sensor Technology[61]

IST has marketed a gas detector with a solid-state semiconductor sensor to detect toxic and explosive gases in air samples. It is developing detectors for hydrazine fuels.[45]

C. Personal Dosimeters or Badges

Mine Safety Appliances (MSA) Company[55]

Organic vapor dosimeter badges, pocket sized CO indicator and alarm

Interscan Corporation[56]

Toxic gas dosimeters for CO, SO_2, H_2S, NO_2, Cl_2, N_2H_4.

MDA Scientific, Inc.[62]

Personal toxic gas dosimeter and alarm for NO_2, H_2S, HCN, $COCl_2$, N_2H_4; chemically impregnated paper tapes for the above gases; portable spot check detector for hydrazines[46]

3M Company[63]

Badges for monitoring CO, organic vapors, ethylene oxide, mercury vapor, formaldehyde

E.E. du Pont de Nemours & Company [64]

Air monitoring badges for organic vapors; colormetric air monitoring badges for SO_2, NO_2, NH_3, EtO, H_2S, formaldehyde

V. RECOMMENDATIONS:

1. The remote sensing instruments are very expensive, usually involve use of lasers, need trained personnel to operate, and usually can detect one or a few pre-determined species. Therefore, at the present time, the use of remote sensing techniques for the detection of hazardous materials from spills does not appear very promising. Nevertheless, the remote sensing is such an ideal feature that the progress in the development of the remote sensing technology should be closely monitored.

2. Many commercial instruments are capable of detecting hazardous gases quickly and accurately. Although they are designed for specific gases, they can be modified to detect other gases such as hypergol vapors. This approach seems to be easier, more economical, and more practical. A number of companies have been developing such instruments or devices. Such instruments should be tested in the laboratories as well as in the field, and the development and improvement should be continued.

3. One of the important problems involved in the treatment of hazardous material spills is the identification of the hazardous material. Two types of instruments that hold the most promise are non-dispersive infrared spectrometers with a computer-based discrimination system and a portable gas chromatograph/mass spectrometer system.

ACKNOWLEDGEMENTS

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research, and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to participate in a very interesting research project in the past summer at the Rocket Propulsion Laboratory, Edwards AFB, CA. He would like to express his deep appreciation to the Laboratory, especially to Mr Forrest S. Forbes, Chief of the Liquid Propellant Systems Section, and Mr John W. Marshall, Program Manager of Hazards Technology, for their suggestions, discussions, and advice in his research effort. Special thanks are due Mrs Eleanore R. Ficke, Technical Librarian, for her assistance in the reference materials.

REFERENCES

1. D.K. Shaver, E.E. Harton, Jr., R.L. Berkowitz, and T.J. Rudd, Post Accident Procedures for Chemicals and Propellants, Air Force Report No. F04611-80-C-0046, Systems Technology Laboratory, Inc., Arlington, Virginia, July 1, 1981.

2. R.M. Measures, "Analytical Use of Lasers in Remote Sensing", in Analytical Laser Spectroscopy, ed, by N. Omenetto, Vol 50 of Chemical Analysis, ed by P.J. Elving and J.D. Winefordner, John Wiley & Sons, New York, 1979.

3. Laser Monitoring of the Atmosphere, ed. by E.D. Hinkly, Vol. 14 of Topics in Applied Physics, Springer-Verlag, Berlin, 1976.

4. Lasers in Chemistry, ed. by M.A. West, Proceedings of the Conference Held at the Royal Institution, London, May 31 - June 2, 1977, Elsevier Scientific Publishing Company, Amsterdam, 1977.

5. W.F. Herget and W.D. Conner, "Instrumental Sensing of Stationary Source Emissions", Environmental Sci. Technology. 11, 962 (1977).

6. W.F. Herget and J.D. Brasher, "Remote Measurement of Gaseous Pollutant Concentrations Using a Mobile Fourier Transform Interferometer System", Appl. Opt. 18, 3404 (1979).

7. P.L. Hanst, A.S. Lefohn, and B.W. Gray, Jr., "Detection of Atmospheric Pollutants at Parts-Per-Billion Levels by Infrared Spectrosopy", Appl. Spectrosc 27, 188 (1973).

8. R.L. Byer and M. Garbuny, "Pollution Detection by Absorption Using Mie Scattering and Topographic Targets for Retroflectors", Appl. Opt. 12, 1496 (1973).

9. W. Schnell and G. Fischer, "Carbon Dioxide Laser Absorption Coefficients of Various Air Pollutants", Appl. Opt. 14, 2058 (1975).

10. C.W. Dewey, Jr., "Optoacoustic Spectroscopy", in Impact of Lasers in Spectroscopy, Vol. 49 of Proceedings of the Society of Photo-Optical Instrumentation Engineers, August, 1974.

11. L.B. Kreuzer, "Ultralow Gas Concentration Infrared Absorption Spectroscopy", J. Appl. Phys. 42, 2934 (1971).

12. L.B. Kreuzer, N.D. Kenyon, and C.K.N. Patel, "Air Pollution: Sensitive Detection of Ten Pollutant Gases by Carbon Monoxide and Carbon Dioxide Lasers", Science 177, 347 (1972).

13. A.M. Augus, E.E. Marinero, and M.J. Colles, "Opto-acoustic Spectroscopy with a Visible CW Dye Laser", Opt Commun 14, 223 (1975).

14. J. Reid, J. Shewchun, B.K. Garside, and E.A. Ballik, "High Sensitivity Pollution Detection Employing Tunable Diode Lasers", Appl. Opt. 17, 300 (1978).

15. R.A. Ferguson, "Feasibility of a CW Lidar Technique for Measurement of Plume Opacity", EPA Report No. 650/2-73-037, 1973.

16. E.E. Uthe, "Lidar Evaluation of Smoke and Dust Clouds", Appl. Opt. 20, 1503 (1981).

17. G.E. Schweitzer, "Airborne Remote Sensing", Environ Sci. Technol. 16, 338A (1982).

18. R.M. Schotland, Proc. Third Symp. on Remote Sensing of the Environment, University of Michigan, Ann Arbor, Michigan, 1964.

19. K.W. Rothe, U. Brinkman, and H. Walther, "Application of Tunable Dye Lasers to Air Pollution Detection: Measurements of Atmospheric NO_2 Concentrations by Differential Absorption", Appl. Phys. 3, 115 (1974).

20. W.B. Grant and R.D. Hake, Jr., "Calibrated Remote Measurements of SO_2 and O_3 Using Atmospheric Backscatter", J. Appl. Phys. 46, 3019 (1975).

21. M.S. Shumate, R.T. Menzies, W.B. Grand, and D.S. McDougal, "Laser Absorption Spectrometer: Remote Measurement of Tropospheric Ozone", Appl. Opt. 20, 545 (1981).

22. H. Inaba and T. Kobaysi, "Laser-Raman Radar-Laser-Raman Scattering Methods for Remote Detection and Analysis of Atmospheric Pollution", Opto-Electron. 4, 101 (1972).

23. T. Kobayasi and H. Inaba, "Spectroscopic Detection of SO_2 and $C0_2$ Molecules in Polluted Atmosphere by Laser-Raman Technique", Appl. Phys. Lett. 17, 139 (1970).

24. S. Nakahara, K. Ito, S. Ito, A. Fuke, S. Komatsu, H. Inaba, and T. Kobayasi, "Detection of Sulphur Dioxide in Stack Plumes by Laser Raman Radar", Opto-Electron. 4, 169 (1972).

25. T. Hirschfeld, E.R. Schildkraut, H. Tannenbaum, and D. Tannenbaum, "Remote Spectroscopic Analysis of ppm-Level Air Pollutants by Raman Spectroscopy", Appl. Phys. Lett. 22, 38 (1973).

26. D.A. Leonard, Laser Raman Gas Diagnostics, ed. by M. Lapp and C.M. Penney, Plenum Press, New York, 1974.

27. C.M. Penny, Laser Raman Gas Diagnostics, ed. by M. Lapp and C.M. Penney, Plenum Press, New York, 1974.

28. M.R. Bowman, A.J. Gibson, M.C.W. Sandford, "Atmospheric Sodium Measured by Tuned Laser Radar", Nature 221, 456 (1969).

29. H. Kildal and R.L. Byer, "Comparison of Laser Methods for the Remote Detection of Atmospheric Pollutants", Proceedings of the IEEE 59, 1644 (1971)

30. R.M. Measures and G. Pilon, "A Study of Tunable Laser Techniques for Remote Mapping of Specific Gaseous Constituents of the Atmosphere", Opto-Electron. 4, 141 (1972).

31. A.W. Tucker, A.B. Peterson, and M. Birnbaum, "Fluorescence Determination of Atmospheric NO and NO_2", Appl. Opt. 12, 2036 (1973).

32. A. Schultz, H.W. Cruse and R.N. Zare, "Laser-Induced Fluorescence: A Method to Measure the Internal State Distribution of Reaction Products", J. Chem Phys 57, 1354 (1972).

33. Laser Spectroscopy of Atoms and Molecules, ed. by H. Walther, Springer-Verlag, Berlin, 1976.

34. R.T. Menzies and R.K. Seals, Jr., "Ozone Monitoring with an Infrared Heterodyne Radiometer", Science 197, 1275 (1977).

35. R.T. Menzies, "Remote Detection of SO_2 and CO_2 with a Heterodyne Radiometer", Appl. Phys. Lett. 22, 592 (1973).

36. R.T. Menzies and M.S. Shumate, "Air Pollution: Remote Detection of Several Pollutant Gases with a Laser Heterodyne Radiometer", Science 184, 570 (1972).

37. C.J. Koh, H.D. Pandey, B.J. Rye, E.L. Thomas, and L.W. Winchester, Jr., "An Infrared DAS Remote Sensing System Using Heterodyne Detection", Lasers in Chemistry, ed. by M.A. West, Proc. of the Conf. Held at the Royal Institution, London, May 31 - June 2, 1977, Elsevier Scientific Publishing Company, Amsterdam, 1977.

38. R.T. Menzies, "Remote Sensing with Infrared Heterodyne Radiometers", Opto-Electron. 4, 179 (1972).

39. P.R. Griffiths, "Infrared Emission Spectroscopy. A Tool for Remote Quantitative Analysis", Spectroscopy, Laboratory Instrumentation, Series II, Vol II, Intern. Scient. Comm., 1977.

40. M. Flanagan, "XM21 Remote Sensing Chemical Agent Alarm: Phase I Scientific and Technical Report", Contract DAAK11-79-C-0051, June 1981.

41. H.W. Prengle, Jr., C.A. Morgan, C.S. Fang, L.K. Huang, P. Campani, and W.W. Wu, "Infrared Sensing and Determination of Pollutants in Gas Plumes", Environ. Sci. Technol. 7, 417 (1973).

42. H. Walter, Jr. and D. Flanigan, "Detection of Atmospheric Pollutants: A Correlation Technique", Appl. Opt 14, 1423 (1975).

43. P.E. Wilkins, "Field Use of IR Spectroscopy for Pollutant Detection", in Analysis Instrumentation, Vol 10, Proceedings of the 18th Annual ISA Analysis Instrumentation Symposium: May 3-5, 1972. ed. by R.L. Chapman, G.A. McNeill, and A.M. Bartz, Instrument Society of America, Pittsburg.

44. R. Dick, "Breadboard Gas Filter Correlation Spectrometer for Atmospheric Measurement of Hydrazines and Nitrogen Dioxide", Rpt No. TR-79-281, Barringer Research Ltd., Rexdale, Ontario, Oct 1981.

45. G.L. Loper, "Hypergol Vapor Detection (HVD) Technical Interchange", Rpt. No. TOR-0081 (6447)-1, Aerospace Corporation, Los Angeles, CA, August 1981.

46. L.J. Luskus, "Measuring Hydrazine Vapor Concentrations in Air", 1981 JANNAF Safety & Environmental Protection Subcommittee Meeting, Kennedy Space Center, Florida, Nov 17-20, 1981. CPIA Publication 348, Nov. 1981.

47. Infrared Industries, Incorporated, Western Division, Instrumentation Group, P.O. Box 989, Santa Barbara, CA 93102. Phone: (805) 684-4181.

48. GOW-MAC Instrument Company, P.O. Box 32, Bound Brook, NJ 08805. Phone: (201) 560-0600 or Western Scientific Association, P.O. Box 698, Danville, CA 94526. Phone: (415) 837-4622.

49. Meloy Laboratories, Inc./Columbia Scientific Industries Corp., P.O. Box 9908, Austin, TX 78766. Phone: (800) 531-5003.

50. The Foxboro Company, Foxboro Analytical, P.O. Box 5449, South Norwalk, CT 06856. Phone: (203) 853-1616 or Flair Business Park, Suite 100, 9420 Telstar Ave., El Monte, CA 91731. Phone: (213) 443-4191.

51. CVC Products, Inc., 525 Lee Rd., P.O. Box 1886, Rochester, NY 14603 or 806 Manhattan Beach Blvd; Suite 208, Manhatten Beach, CA 90266. Phone: (213) 376-0407.

52. Beckman Instruments, Inc., Process Instruments Division, 2500 Harbor Blvd., Fullerton, CA 92634. Phone: (714) 773-8175.

53. Aero Chem Laboratories, Inc., P.O. Box 12, Princeton, NJ 08540.

54. The Aerospace Corporation, P.O. Box 92957, Los Angeles, CA 90009. Phone: (213) 648-7418.

55. Mine Safety Appliances (MSA) Company, 600 Penn Center Blvd., Pittsburg, PA 15235. Phone: (412) 273-5000 or 7100 Fair Ave., North Hollywood, CA 91605. Phone: (213) 875-2905.

56. Interscan Corporation, 21700 Nordhoff St., Chatsworth, CA 91311. Phone: (213) 882-2331.

57. The Mogul Corporation, P.O. Box 200, Chagrin Falls, OH 44022. Phone: (216) 247-5000.

58. Hach Company, 1800 Wyatt Dr., Suite #1, Santa Clara, CA 95054. Phone: z (408) 980-1120.

59. Bacharach Instrument Co., 301 Alpha Dr., Pittsburg, PA 15238. Phone: (412) 782-3500.

60. Energetics Sciences, Inc., Division of Bector Dickinson & Company, 6 Skyline Dr., Hawthorne, NY 10532. Phone: (914) 592-3010.

61. International Sensor Technology, 3201 South Halladay St., Santa Ana, CA 92705. Phone: (714) 546-0672.

62. MDA Scientific, Inc., 1815 Elmdale Ave., Glenview, IL 60025. Phone: (800) 323-2000.

63. 3M Company. Occupational Health and Safety Products Division, 220-7W, 3M Center, St. Paul, MN 55101. Phone: (612) 733-8831.

64. E.I. du Pont de Nemours & Company. Applied Technology Division Brandywine Bldg. 4300, Wilmington, DE 19898. Phone (302) 774-2421.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

TENTATIVE IDENTIFICATION AWARENESS STAGES LEADING TO FAILURE OF VIGILANCE

Prepared by:	Dr. Michael L. Lobb
Academic Rank:	Assistant Professor
Department and University:	Graduate School of Social Work The University of Texas at Arlington
Research Location:	Crew Performance Branch USAF School of Aerospace Medicine Brooks AFB, Texas
USAF Research Colleague:	Dr. John Courtright
Date:	September 27, 1982
Contract No:	F49620-82-C-0035

TENTATIVE IDENTIFICATION AWARENESS STAGES LEADING TO FAILURE OF VIGILANCE

by

Michael L. Lobb

ABSTRACT

Vigilance, or more specifically sustained attention, has been studied by various electrophysiological and behavioral methods. Electrophysiological measures, typically EEG, are equivocal indices of sustained attention, whereas the behavioral measure of the missed signal is convincing. A measure of sustained attention more reliable than EEG is clearly needed. In the present paper, electrooculographic (EOG) measures are taken along with simultaneous video taping of EOG, EEG, and actual eye movements during a temporal auditory discrimination task. The computer record of these measures is then analyzed for sequences of eye movements and EEG which fall into discernable patterns.

The results suffice to tentatively identify at least five substages of sustained attention: (1) involuntary blinks, (2) voluntary long closing duration blinks, (3) long closure duration with SEM precursors, (4) long closure durations with alpha and half-wave SEM, and (5) long closure durations with full wave SEM and alpha fading during the closure. These substages appear in episodes and roughly describe the progression from sustained attention to missed signals. In addition to these substages saccadic reorientation eye movement(s) may follow the termination of a substage 4 or 5 occurrence. These findings, although tentative, may offer a reliable means of measuring the central state of sustained attention. Furthermore, transitions within and between episodes of sustained attention do not follow an all or none law, but rather are progressive and orderly.

I. INTRODUCTION

A review of the vigilance literature, as compiled by Mackie[1] indicates that vigilance research operates from the tacit assumption that vigilance is an all or none function. Either the subject is vigilant or not, and the lower boundary is the missed signal. Numerous attempts have sought to show electrophysiological correlates with equivocal results although some potential for fruitful research is indicated Gale[2]. Recently, however, Stern, Walrath, and Goldstein[3] have reported the endogenous eye blink to be an indicant at central state sensitive to time-on-task manipulations. This report introduces the possibility of measuring central state changes leading to the failure and vigilance through electro-oculographic (EOG) techniques.

The study of eye movements must be given a special place in any theoretical view of macro-behavior such as vigilance. Fine-grained eye movements either in response to or in the absence of field stimuli represents the actions of a CNS managed system preparing to receive information prior to the steps of information processing or response output. Measures of oculomotor activity antecede measures of decision time, memory retrieval, or reaction time - in terms of the S-O-R macro-model the oculomotor system falls between the "S" and the "O", chaining notwithstanding. The critical difference is that fine-grained sensory orientation movements antecedent to higher mental processing and macro-response making are far more likely to be patterned. And the state of sensory preparedness to receive information, as indicated by patterns of fine-grained orientation, may be a direct reflection of central state.

The study of fine-grained sensory orientation patterns has not received systematic attention in the literature of American psychology as it

has in the European study of ethology. As can be seen in most textbooks of ethology, the study of motor patterns is typically directed toward lower animal forms in which behavioral systems are simpler and the occurrence of motor patterns more regular. A cursory glance at recent pilot data on eye movements indicates that the patterns are sufficiently simple and regular to allow a temporal analysis of sequential behaviors such as that described by Lobb[4].

II. OBJECTIVES OF THE RESEARCH EFFORT

The goal of this effort is to simultaneously record EOG, EEG, and video taped eye movements for a preliminary examination of patterns. The data set is to be subsequentially transcribed and used for a temporal analysis of sequential behavior, a task beyond the scope of this report.

III. METHODS

Seven subjects were selected from the available staff of the Crew Performance Branch, USAF School of Aerospace Medicine, Brooks AFB, Texas. Each subject was informed of the procedures, offered the opportunity to withdraw at any time, and asked to sign an informed consent statement (copy attached). Electroencephalographic (EEG) electrodes were attached with collodion at P3 and O1 with reference at A2 according to the Ten Twenty Electrode System (one subject declined the EEG electrodes). Electro-oculographic (EOG) electrodes were attached above and below the right eye to measure lid and/or eye deflections in the vertical plane. A pair of EOG electrodes were also attached one laterally to the side of the head behind each eye so as to measure horizontal deflections. The EOG reference electrode was attached at the center of the forehead and DC

recordings made below 30 hz. Video tapes were made of the right eye, EEG, and EOG by means of a split screen.

Following electrode preparations each subject was seated in a reclining chair, instructed in the use of the response apparatus (a hand held button), and ask to put on a set of ear phones. When the Beckman twelve channel recorder and the video tapes recorder were indicating appropriate signals, the subject was instructed to look at a fixation point in the center of his visual field, and presented with progressive relaxation instructions via an audio tape. The EEG, EOG, and video were monitored until the conjoint occurrence of alpha waves, long duration lid closures (greater than 2 seconds), and slow eye movements (SEM). At that point the progressive relaxation tape was discontinued and a temporal auditory task presented via a tape recorder.

The temporal auditory task consisted of 200 or 400 millisecond tones presented on a fixed interval or variable interval schedule as described by Stern[4]. The subject was instructed to push the response button once for a short tone and twice for a long tone. A 16.7 millisecond, stop frame video editor was used to study the data.

The EEG channel permitted identification at alpha and non-alpha periods. A visual study at the horizontal (HEOG) and vertical (VEOG) channels along with the simultaneous video record of eye lid and ball movement permitted the identification of the following coding system in which all motion is judged by changes in position between 16.7 millisecond time-frames:

1. Lid Up (LU): Lid up and not moving in the video picture with the VEOG pen indicating no further deflection. LU may be seen with the lid in the upper orbit of the pupil, above the pupil,

or above the iris; but the chief identifying feature is that the lid is up (above the mid line of the pupil) and static (not in motion).

2. Lid Above Pupil (LAP): Lid in downward motion from a LU position or upward from a LOP motion component. This is the motion component as the lid approaches the pupil indicated by a positive deaccelerating component in the VEOC.
3. Lid On Pupil (LOP): Lid in motion up or down as it passes over the pupil.
4. Lid Below Pupil (LBP): Lid in upward or downward motion between the lower orbit at the pupil and the lid down static position - all motion below the pupil.
5. Slow Eye Movement (SEM): Slow, sinusoidal movements on the HEOG at 1/4 to 1/2 cycle per second.
6. Static (ST): No indicated deflection in HEOG.
7. Up Saccade (US): High velocity upward eye movement seen on the video record and indicated in the VEOG square wave deflection often obscured by a blink.
8. Down Saccade (DS): High velocity downward eye movement seen in the video record and indicated in the VEOG square wave deflection.
9. Left Saccade (LS): High velocity leftward eye movement seen in the video record and indicated by a square wave deflection in the HEOG.
10. Right Saccade (RS): High velocity rightward eye movement seen in the video record and indicated by a square wave deflection in the HEOG.

11. Return Saccade (SR): High velocity left or right up or down, saccadic motion during lid opening which results in returning the eye to the approximate fixation point occupied before lid closure.

12. Realignment Saccade (RA): High velocity left or right, up or down saccadic motion that corrects for error (overshoot) in the return saccade.

IV. RESULTS

The lid closing sequences (LU → LAP → LOP → LBP → LD) vary in duration in contrast to the lid opening sequence (LD → LBP → LOP → LAP → LU) which is relatively stable in duration. Variations in the lid closing sequences are associated with extended LAP components whereas variations in closure duration are associated with extended LD components.

The occurrence of a series of involuntary blinks during undefined EEG activity and without variations in the HEOG describe substage 1. This substage is shown in Figure 1 in which time progresses from left to right. In the involuntary blink sequence the LAP component is reduced in the closing portion giving the VEOG record a square wave effect, whereas in the opening component the LAP is extended producing a positive deacceleration in the VEOG record. Notice that the record in Figure 1 is episodic; that is, the same involuntary blink sequences reoccur forming the involuntary blink episode characteristic of substage 1.

Substage 2 is depicted in Figure 2. Here LAP is extended in both the closing and opening components, and deflections in the HEOG record occur within the closing LAP to opening LAP interval. As the duration of LD is extended the deflections in HEOG appear as precursors of SEM an example of

which can be seen in the first closure shown in Figure 4. In substage 2 the EEG record is undefined.

In substage 3 as shown in Figure 3 the LAP, LOP, and LBP components are extended during the long closing duration component giving the VEOG record a diagonal component, while the opening component is unchanged. Although the EEG record is still undefined, wave forms very like alpha appear within the second closure. The HEOG record shows clear precursors of SEM within closure intervals, although these precursors do not yet have the sinusoidal appearance of fully developed SEM.

In Figure 4 the LD component is extended producing a long duration closure characteristic of substage 4. Alpha is clearly defined and distributed within the LCD interval. SEM appears in half-wave sinusoidal form and is also distributed within the LDC interval. Notice also the occurrence of pairs of left and right saccades with a 200 to 300 MS latency between deflections. These saccades follow the opening component at a latency less than the LDC interval. Notice also that these pairs of saccades describe deflections in a right or left visual field, but not both, and that the second saccade returns the eye to the fixation point from which the first saccade deflected. This substage is best described as a simultaneous alpha, LDC, 1/2 wave SEM followed by pairs of saccades. This sequence is clearly episodic even though substage 3 SEM precursors antecede and follow the episode.

In Figure 5 the LD component of closure is further extended, alpha appears early in the interval and fades before the opening component. This fifth substage is characterized by alpha and SEM activity which do not fill the entire LDC interval. Although not shown in Figure 5, the fifth substage may be episodic with closure durations which exceed the

30 seconds which can be shown on a page. Pairs of saccades also commonly follow the eye opening component. Fully sinusoidal SEM is seen in Figure 6.

Although these substages are progressives as a function of time on task, transition between substages are observed.

V. DISCUSSION

The preliminary results of this study indicate the existence of orderly substages in the central state of sustained attention. The duration of the LAP component differentiates the voluntary from the involuntary blink and hence substage 1 from substage 2. This differentiation is based singularly on the lid closing segment in which the LAP duration increases. In substage 3 the LOP and LBP components increase in duration, and precursors of SEM appear. An increase in LD duration describes substage 4 in which alpha appears along with 1/2 wave SEM. In substage 5 alpha fades during the extended LD component and SEM is fully developed. The transition from substage 4 or 5 to substage 1 may be accompanied by saccadic reorientation eye movements in a left or right field but not both. Although the precise identification of substages awaits completion of data transcription and analysis, these preliminary results strongly suggest that the loss of attention is not an all or none function, but rather an orderly progression of substages.

VI. RECOMMENDATIONS

The transcription and analysis of data continues beyond the scope of the present project. Each 16.7 MS time-frame is inspected and the EEG, VEOG, and HEOG events transcribed to create a three channel data set for

computer analysis. The analysis envolves counting sequential events between and within channels. Between channels events initiated simultaneously are counted (the zero order transition), and then, separately for each event, all first, second, and third order transitions are counted both antecedent and consequent to each event. The within channel analysis is similar except that the zero order transition does not exist within a channel. For each and every transition the latency is counted as is the duration of the event. The criteria for concluding that any set of events is sequential is twofold: first the serial orderliness must be significantly higher and second, the variance of the latencies must be significantly lower. Inferential statistical comparisons are made by application of a within subjects (mean deviation) t-test in which average conditional probability of a sequence of any set of events during a given substage is compared to the average conditional probability of the same set of events during a randomly selected equal time interval. A similar comparison is made for the variance of transitional latencies. Once sequences have been identified then the data is reanalyzed for episodes or series of the same sequences.

The basic software for this analysis has been developed at the University of Texas at Arlington and computer time on the IBM 4341 has been made available by the University. However, a video editor for data transcription and additional software unique to this project is needed. Support for these requirements will be sought through an AFOSR minigrant. I recommend that support be given to this endeavor. Transcription and analysis of the data collected will take about five months, and salary support for this project will be forthcoming from the University of Texas at Arlington.

Acknowledgement

I wish to thank Drs. Bryce Hartman and John Courtright, and the staff of the Crew Performance Branch of the USAF School of Aerospace Medicine at Brooks AFB, Texas for their professional support and friendly assistance. I also wish to thank Dr. Douglas Eddy of Trinity University, San Antonio, Texas for his encouragement and logistical support. Finally, I want to express a special sense of gratitude to Dr. John Stern, Washington University, Saint Louis, Missouri for his guidance and collegial affiliation.

REFERENCES

1. R. R. Mackie, Ed., Vigilance, Theory, Operational Performance, and Physiological Correlates, Plenum Press, New York, 1977.

2. A. Gale, in Vigilance Theory, Operational Performance, and Physiological Correlates, Chapter 15, R. R. Mackie, Ed., Plenum Press, New York, 1977.

3. J. A. Stern, L. C. Walrath, and R. Goldstein, "The Endogenous Eyeblink: An Indicant of Cognitive Activity and State," A preprint from Washington University Behavior Research Laboratory, 1420 Grattan St., St. Louis, Missouri 63104.

4. J. A. Stern, Washington University, St. Louis, Missouri 63104. Personal communication.

5. M. L. Lobb, "A Temporal Analysis of Sequential Courtship Behaviors of Sunfish Genus Lepomis," manuscript submitted for publication, 1982.

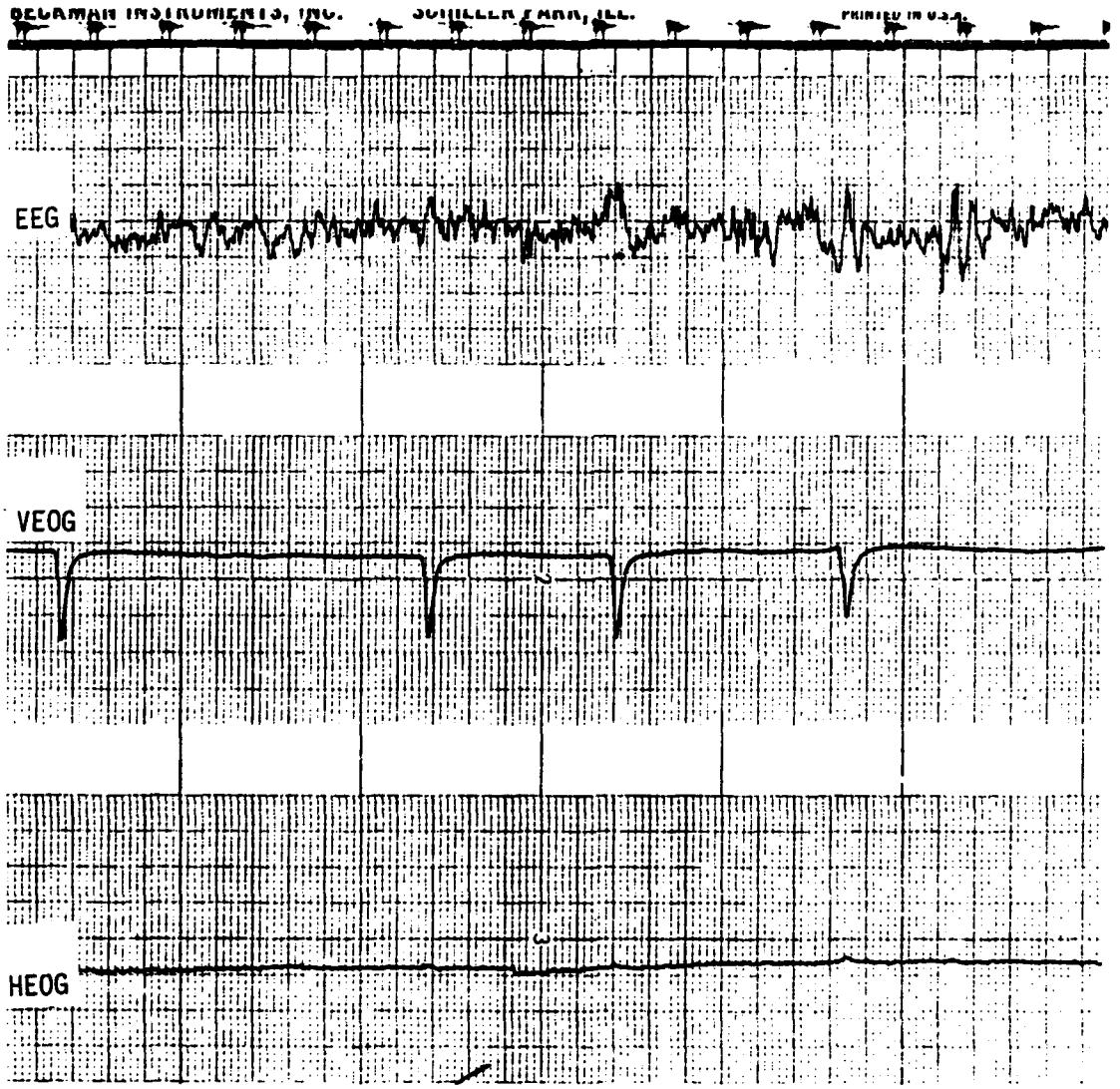


FIGURE 1
SUBSTAGE 1

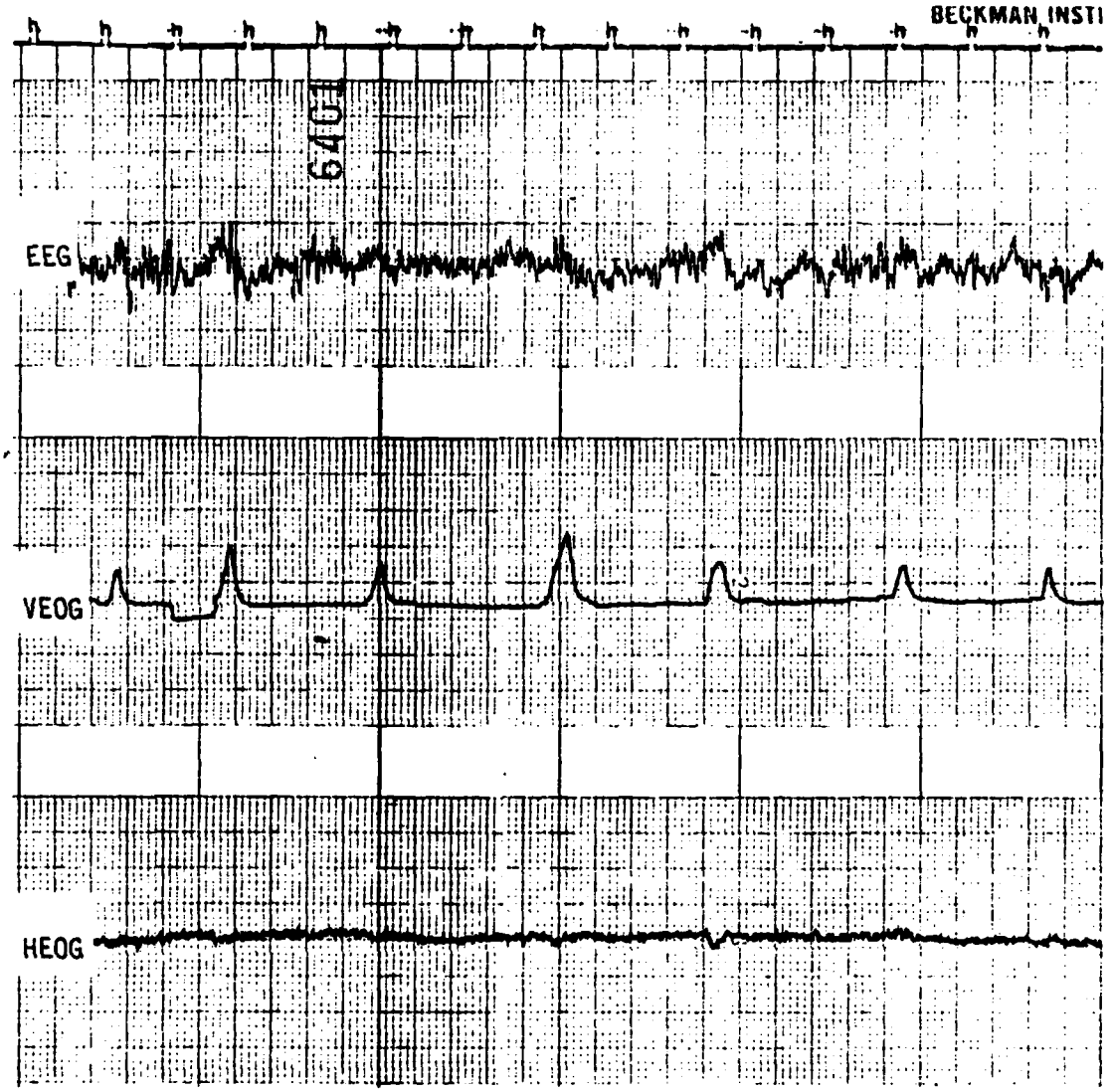


FIGURE 2
SUBSTAGE 2

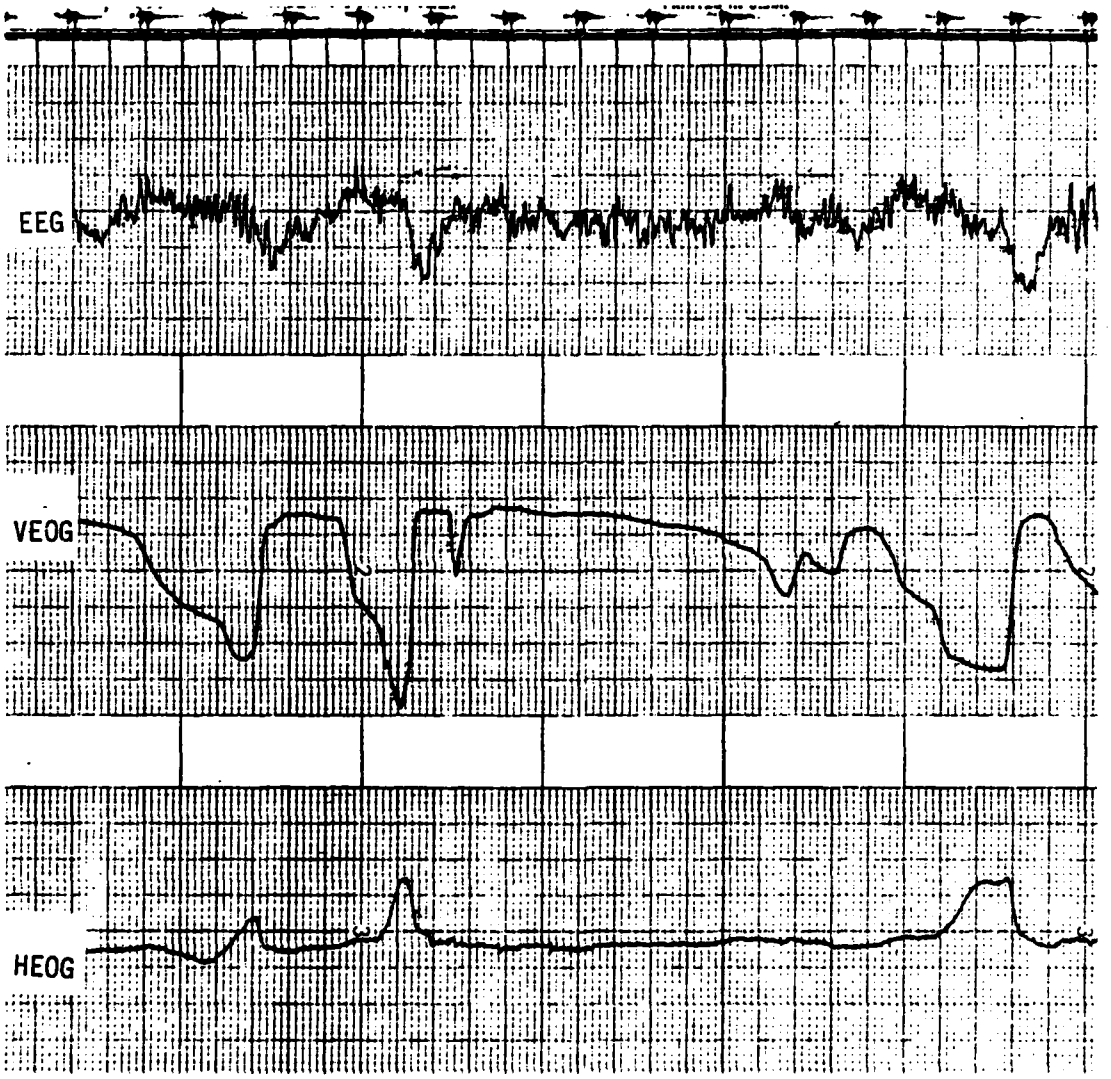


FIGURE 3
SUBSTAGE 3

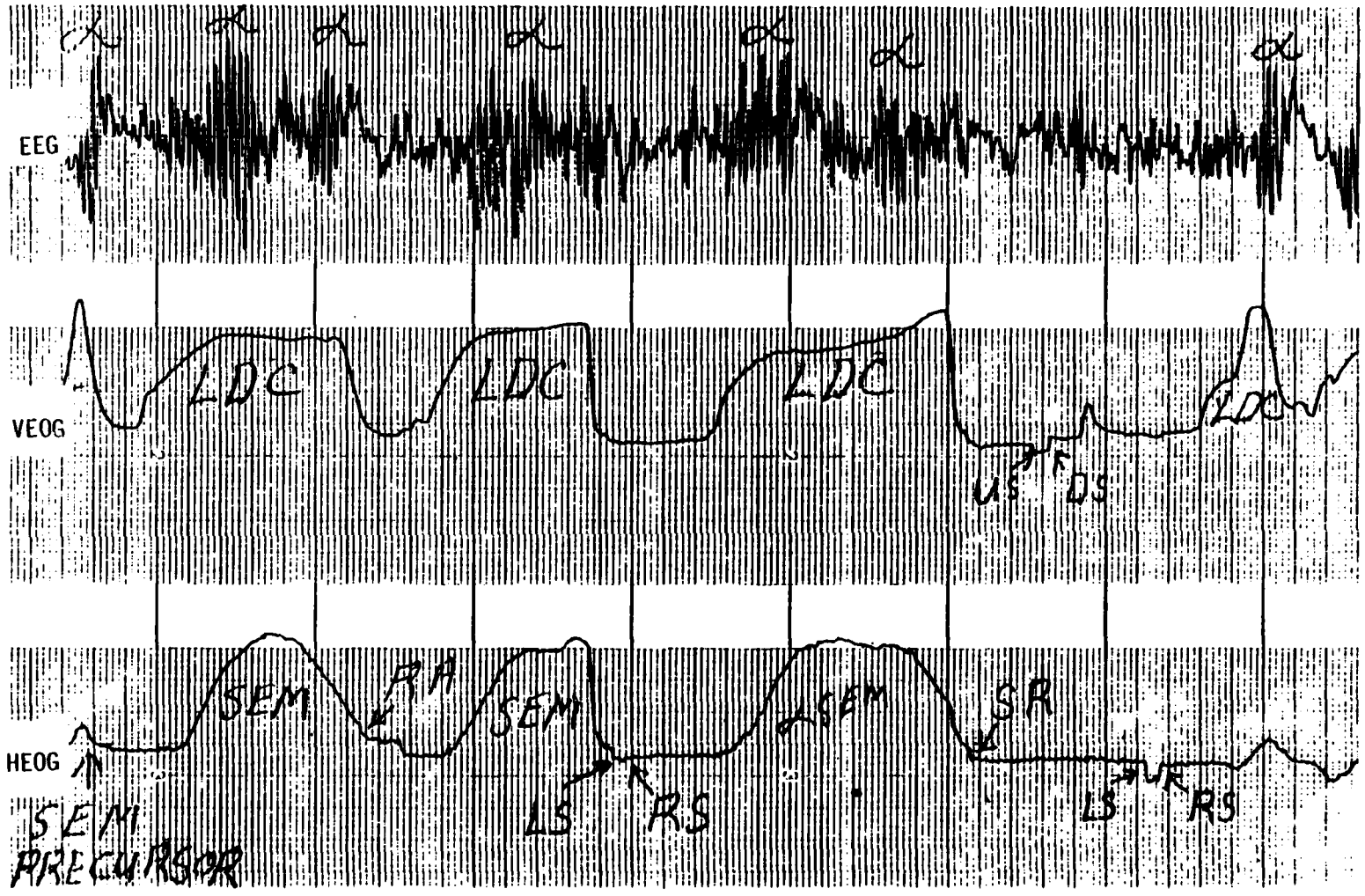


FIGURE 4
SUBSTAGE 4

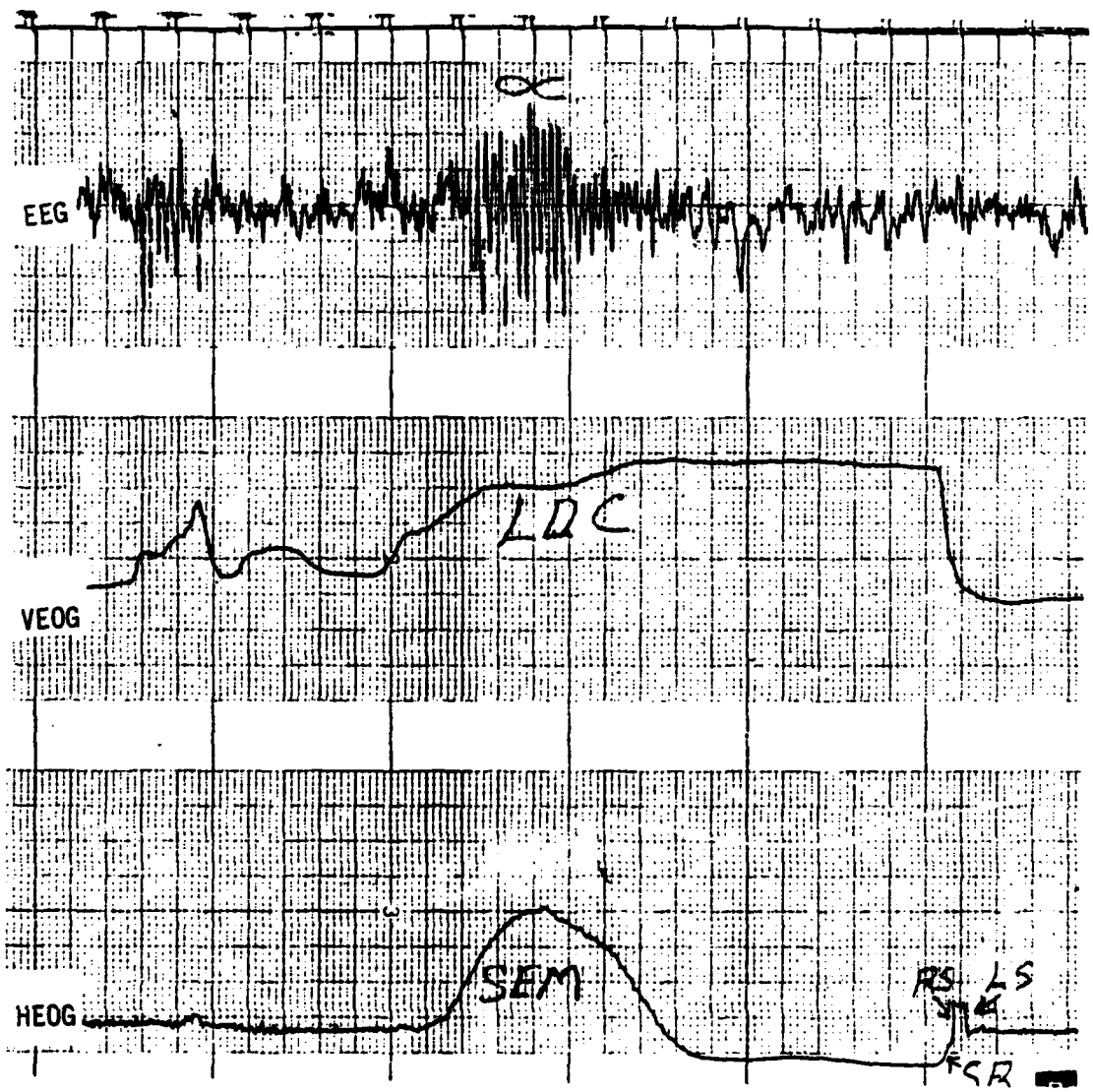


FIGURE 5
SUBSTAGE 5

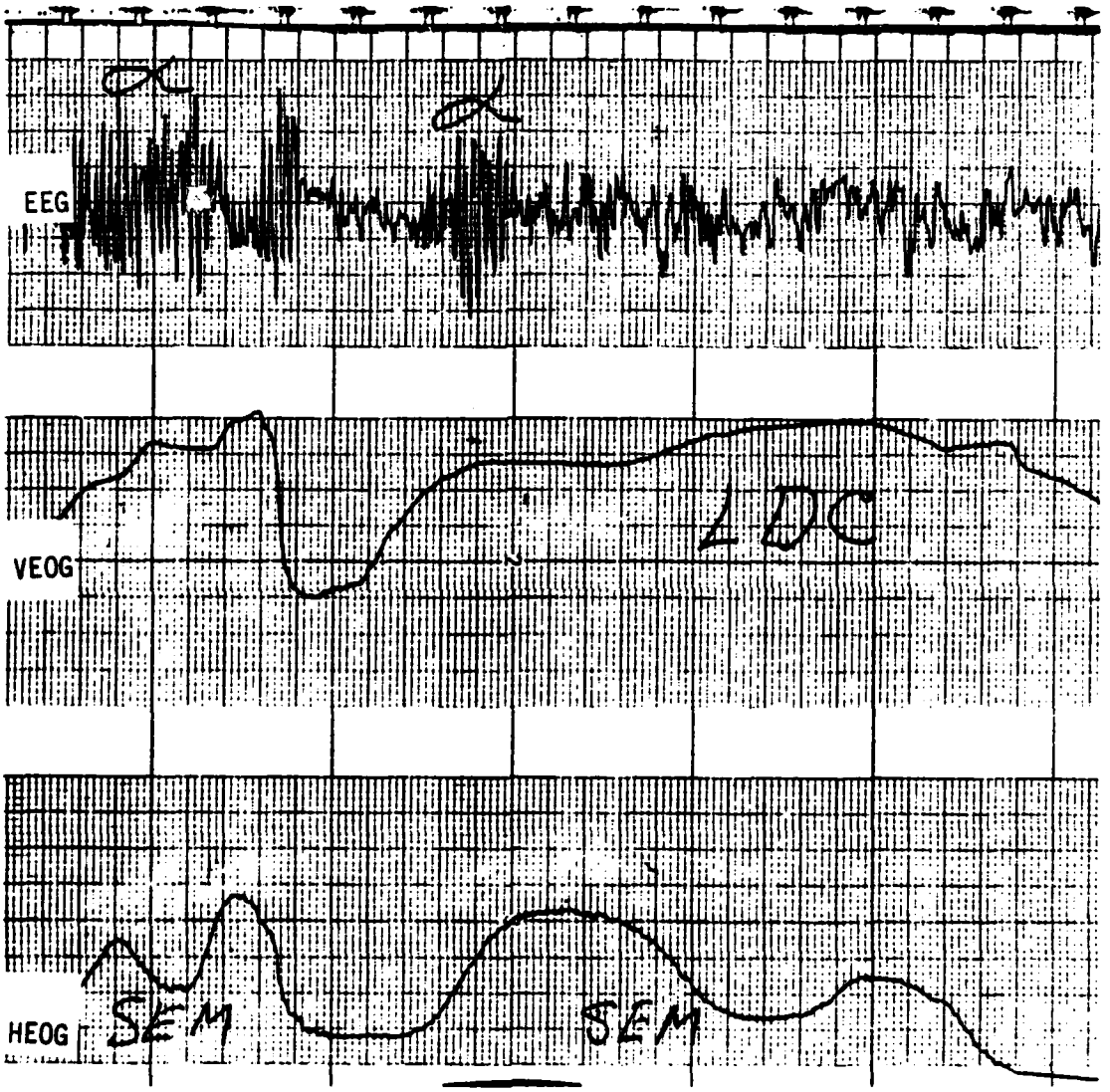


FIGURE 6
SUBSTAGE 5

INFORMED CONSENT JULY-AUGUST 1982

I, _____, having full capacity to consent, do hereby volunteer to participate in a research study entitled "Sequential Eye Movements During Stage I Sleep Transition. A preliminary investigation under the direction of Michael Lobb, Ph.D. The implementation of my voluntary participants, the nature, duration, and purpose; the initiation and means by which the study is to be conducted; and the inconveniences and hazards which may resonably be expected have been explained to me by Michael Lobb, and well described on the reverse side of this consent form which I have read and initialed. I have been given an opportunity to ask questions about this research and received answers to my full and complete satisfaction. I understand that I may withdraw at any time and revoke my consent without prejudice. I understand the study will be completed in strict confidence.

Subject: _____ Date: _____

PROCEDURES, ADVANTAGES, AND DISADVANTAGES

1. Three scalp electrodes and five facial electrodes will be attached by means of colodian in the case of scalp electrodes and by adhesive tape in the case of EOG electrodes. Small amounts of alcohol will be used to clean these areas and the removal of hairs about 1 1/2 centimeters in area. Dead skin cells will be removed by the use of an abrasive and the use of alcohol.

2. Each subject will be trained in the Jacobson Progressive Relaxation Procedures which will result in an enhanced ability to relax.

3. Each session will run three hours including preparation time and data collection. Following data collection the results will be presented to each participant.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

ORDER RELEASE IN AN MRP ENVIRONMENT

Prepared by:	Dr. Deborah J. Medeiros
Academic Rank:	Assistant Professor
Department and University	Department of Industrial and Management Systems Engineering, The Pennsylvania State University
Research Location:	Air Force Wright Aeronautical Laboratories, Materials Laboratory, Manufacturing Technology Division, Computer Integrated Manufacturing Branch
USAF Research Colleague:	Richard Mayer
Date:	September 10, 1982
Contract No.	F49620-82-C-0035

ORDER RELEASE IN AN MRP ENVIRONMENT

by

Deborah J. Medeiros

ABSTRACT

To date work in Manufacturing Resource Planning (MRPII) has concentrated on the development of order release and due dates based on fixed lead times, a master production schedule, and a capacity plan. This paper discusses an extension of the order release procedure which controls work in process inventory through staged release of batches of orders to the factory floor. The composition of each batch of orders varies depending on queue lengths at key operations. The objective of the procedure is to reduce overall flowtime.

The procedure uses a macro level simulation model to assess the effects of changing the order release times. The simulation is written in $IDEF_2$, a dynamics modeling methodology developed for the ICAM Program. The use of $IDEF_2$ in this environment is discussed. Suggestions for extension of the research are provided.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

THE MEASUREMENT OF ION-MOLECULE REACTION RATE COEFFICIENTS

Prepared by:	Dr. Thomas M. Miller
Academic Rank:	Associate Professor
Department and University:	Department of Physics and Astronomy University of Oklahoma
Research Location:	Air Force Geophysics Laboratory, Aeronomy Division
USAF Research Colleague:	Dr. John F. Paulson
Date:	August 9, 1982
Contract No:	F49620-82-C-0035

This report was prepared with Rodney E. Wetterskog, graduate student, University of Oklahoma.

THE MEASUREMENT OF ION-MOLECULE REACTION RATE COEFFICIENTS

by

Thomas M. Miller and Rodney E. Wetterskog

ABSTRACT

Rate coefficients have been measured for C^+, CO^+, and CO_2^+ reactions with O_2 over a temperature range of 90-450 K using a selected-ion flow-tube (SIFT) apparatus at the Air Force Geophysics Laboratory. Charge transfer is observed to take place for $CO^+ + O_2$ and $CO_2^+ + O_2$, but is not allowed energetically for $C^+ + O_2$. Instead, we observe $C^+ + O_2 \rightarrow O^+ + CO$ and $C^+ + O_2 \rightarrow CO^+ + O$ occurring. The $C^+ + O_2$ reaction is found to proceed at its gas kinetic rate of about 9×10^{-10} cm^3/s, independent of temperature for 90-450 K. The rate coefficients for both the $CO^+ + O_2$ and $CO_2^+ + O_2$ reactions are decreasing functions of temperature in this range. The ion flow velocity in the SIFT apparatus has been determined versus gas pressure and temperature using a time-of-flight technique. These results, combined with our helium flow measurements, have significant implications for low-temperature reaction experiments with flow-tube reactors.

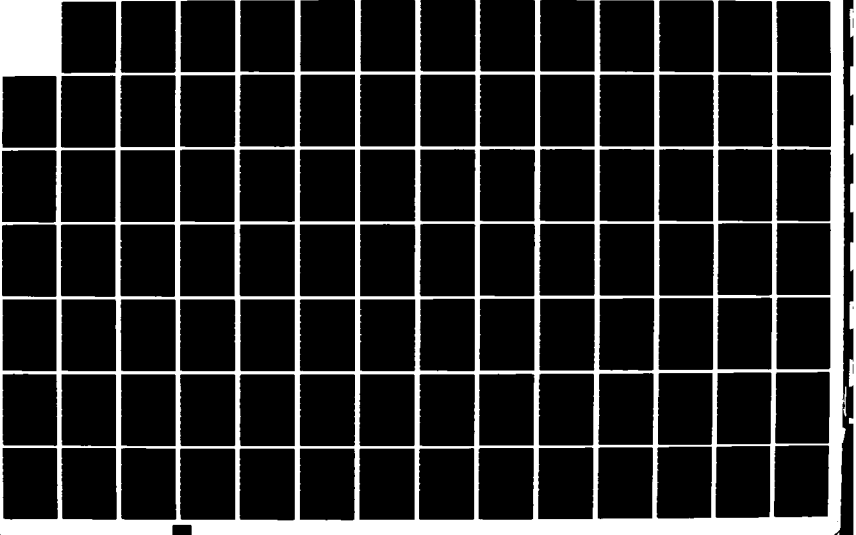
I. INTRODUCTION:

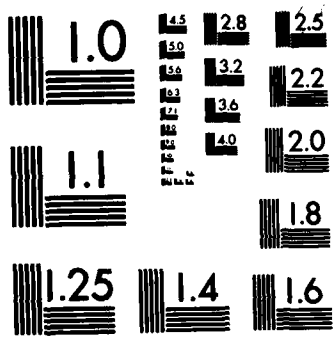
The Aeronomy Division at the Air Force Geophysics Laboratory has long been active in research on the composition and chemistry of the earth's upper atmosphere. This research involves both atmospheric sampling and laboratory studies of atmospheric reactions. The laboratory research centers on charged-particle interactions primarily because of the importance of military communications through (or reflected by) the partially ionized gas of the earth's atmosphere. The particular laboratory we were working with (AFGL/LKB) is currently studying atmospheric ion reactions with neutral constituents, and the photodissociation of ionic molecules.

The main interest at present in the AFGL ion-molecule reaction research concerns ion clusters which have been found to dominate the stratospheric ion population. Since there has been very little work done on negative ion reactions the emphasis has been in that direction. (The negative ion reactions tend to be more difficult to study because generally speaking it is easier to make large densities of positive ions than negative ions.) The neutral reactants chosen for these experiments are of course observed (or suspected) atmospheric molecules which may be present in the atmosphere in very small fractional concentrations (mixing ratios $\leq 10^{-6}$) but nevertheless play an important role because of their large dipole moments or low ionization potentials.

At AFGL/LKB a selected-ion flow-tube (SIFT) apparatus was constructed about two years ago to study ion-molecule reactions. An essential feature of the design of the AFGL SIFT apparatus is its ability to operate over a temperature range of 80-500 K, which allows the determination of ion-molecule reaction rate coefficients for

AD-A130 770 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM (1982) 3/11
RESEARCH REPORTS VOLUME... (U) SOUTHEASTERN CENTER FOR
ELECTRICAL ENGINEERING EDUCATION INC S...
UNCLASSIFIED W D PEELE ET AL. OCT 82 AFOSR-TR-83-0614 F/G 5/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

atmospheric temperatures (180-300 K) as well as providing for a wider range necessary for insight into the reaction mechanisms. Much of the work we were involved in during the summer of 1982 was in taking the SIFT apparatus to low and high temperatures for the first time and the solution of various problems encountered. For this purpose relatively straightforward ion-moleucle reactions were chosen, $C^+ + O_2$, $CO^+ + O_2$, and $CO_2^+ + O_2$. Our objective was to study the fate of these reactions over a wide temperature range.

One of us (TMM) has had experience with ion-molecule reaction research at the University of Birmingham (UK) on the original SIFT apparatus and at Georgia Tech with a drift tube apparatus. Both of us have worked with a flow tube reactor at our home institution, the University of Oklahoma.

II. APPARATUS AND EXPERIMENTAL METHOD:

The selected-ion flow-tube method was developed by Adams and Smith[1] at the University of Birmingham (UK) around 1975 and represented a significant improvement over the already quite successful flowing afterglow technique. The SIFT method is still being improved upon as its use spreads.

Briefly, the AFGL SIFT apparatus consists of a 108-cm long, 7.3-cm i.d. tube through which a helium carrier gas is flowed at low pressure (~0.3 torr) and high speed (~10^4 cm/s). Ions are created in an electron-impact ion source, mass analyzed, and a current of ~10^{-10} A of a single desired ion species is injected into the flow tube at low energies in the center of a supersonic jet of the helium carrier gas. The ions quickly thermalize in collisions with the helium gas and flow

with the gas down the tube. The trip requires approximately 10 ms. At the end of this trip some of the ions pass through a small (0.03 mm) aperture leading into a second mass spectrometer maintained at high vacuum. The detected ion count rate is proportional to the original injected current.

In order to determine ion-molecule reaction rate coefficients, a very small concentration of reactant gas is titrated into the flow tube at a point sufficiently downstream from the ion and helium injector that equilibrium flow has been reached. If the primary ions react with the titrated gas, producing a product ion of different e/m, the detected primary ion current I_o is attenuated exponentially by the reaction according to Beers' Law:

$$I = I_o e^{-knt}. \tag{1}$$

The reaction rate coefficient k is calculated from measurement of the fractional decrease $I/1_o$ in the primary ion current by a known concentration of reactant gas n, in a time t:

$$k = \frac{\ln(I_o/I)}{nt}. \tag{2}$$

Experimentally, the small fractional concentration n of the reactant gas is determined from the measured helium density N and the throughput Q_R of the reactant gas compared to the helium throughput Q_{He}. The reaction time t is determined from a direct measurement of the ion velocity v_i and of the distance d over which the reactant gas flows. Thus,

$$k = \frac{\ln(I_o/I)}{(Q_R\ N/\ Q_{He})(d/v_i)}. \tag{3}$$

A computer program, SIFT82, has been written to handle the SIFT data on the laboratory's Hewlett-Packard 2112 computer. The program plots the data, calculates the reaction rate coefficient and related quantities, and calculates and plots the ion product percentages.

An example of the raw data is given in Fig. 1 for the reaction $C^+ + O_2$ at 196 K. The reaction time in this case is 3.84 ms and the reaction distance is 43.5 cm. The direct ionic products of the reaction are O^+ or CO^+ ions. The O_2^+ ions observed are secondary products from the subsequent reactions $O^+ + O_2$ and $CO^+ + O_2$. The reaction rate coefficient is 8.8×10^{-10} cm^3/s.

Three different titrant ports are available to enter reactant gases, at three different distances, permitting any end effects to be accounted for.

Ion-molecule reactions are examined for temperatures other than room temperature by cooling or heating the flow tube and the helium gas.

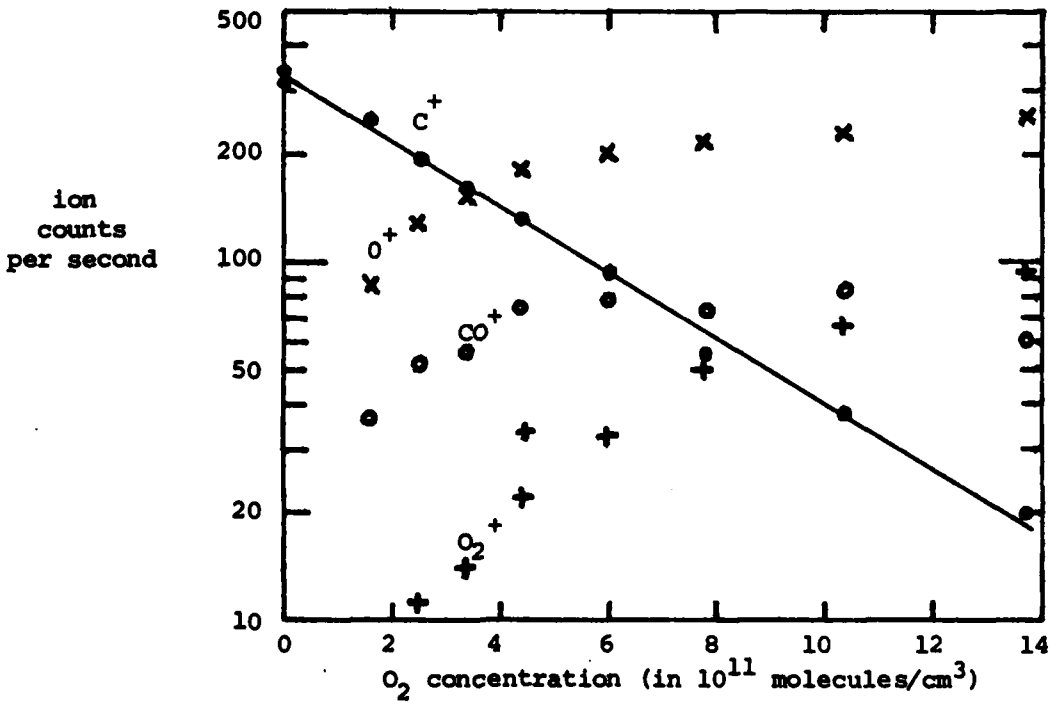


FIG. 1. DATA FOR THE REACTION $C^+ + O_2$ AT 196 K.

III. RESULTS:

A. Ion-Molecule Reactions. The reactions $C^+ + O_2$, $CO^+ + O_2$, and $CO_2^+ + O_2$ have been studied over the temperature range 90-450 K. The results are shown in Figs. 2-4, compared to previous work.

The interaction $C^+ + O_2$ proved to be independent of temperature over the range 90-450 K and essentially equal to the calculated gas kinetic (Langevin) rate coefficient of 1.0×10^{-9} cm^3/s. Simple charge transfer is endoergic in this case. Two reaction channels were observed, however. The first, $C^+ + O_2 \rightarrow O^+ + CO$ is exoergic by 3.635 eV, and is favored (62%) at all temperatures used. Although the neutral CO product is not observed directly, the energetics of the reaction require that it be bound. The second channel observed is $C^+ + O_2 \rightarrow CO^+ + O$, which is exoergic by 3.240 eV.

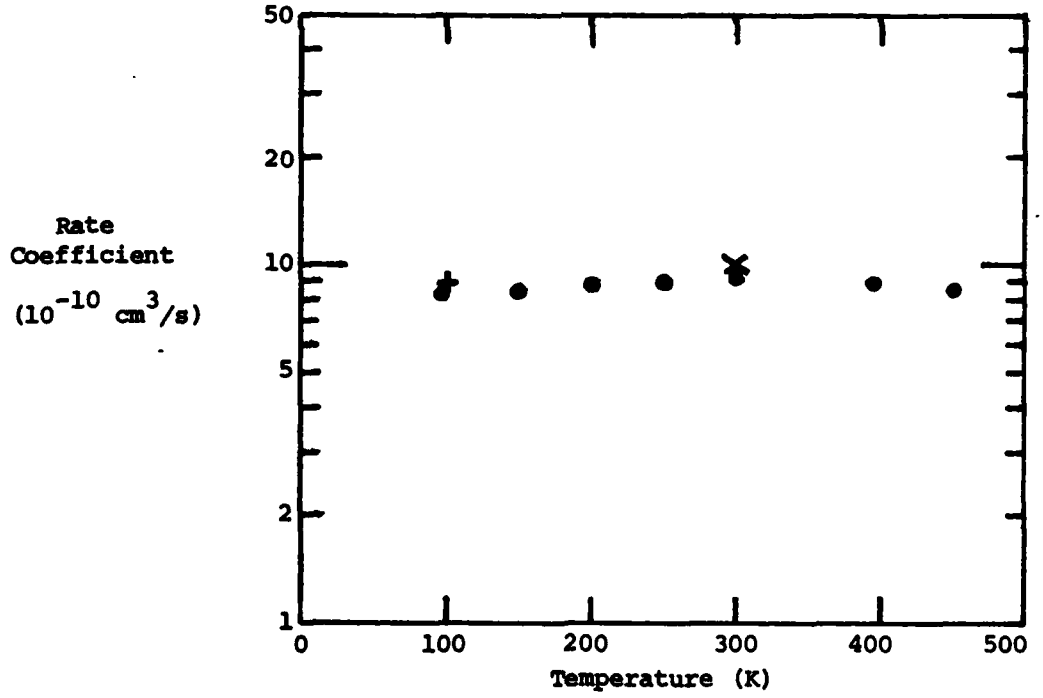


FIG. 2. REACTION RATE COEFFICIENT FOR $C^+ + O_2 \rightarrow O^+ + CO$. The data point ✕ is that of Adams and Smith,[2] and of Rakshit et al.[3] The point + is that of Rakshit et al.[3]

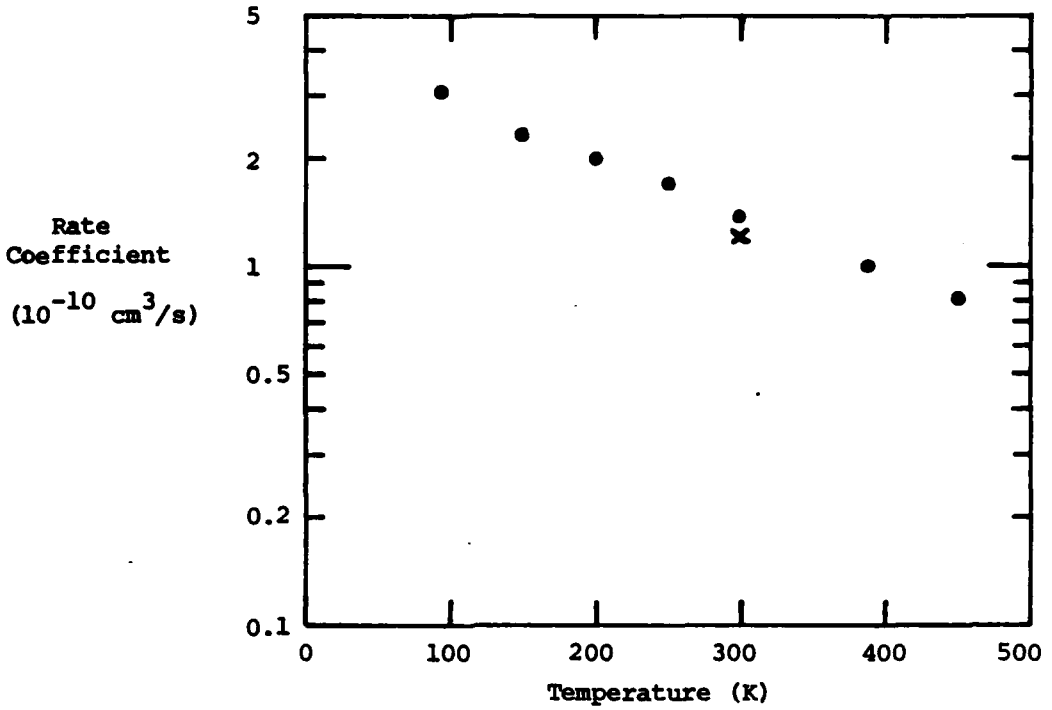


FIG. 3. REACTION RATE COEFFICIENT FOR $CO^+ + O_2 \rightarrow CO + O_2^+$. The data point ✖ is that of Adams et al.[4]

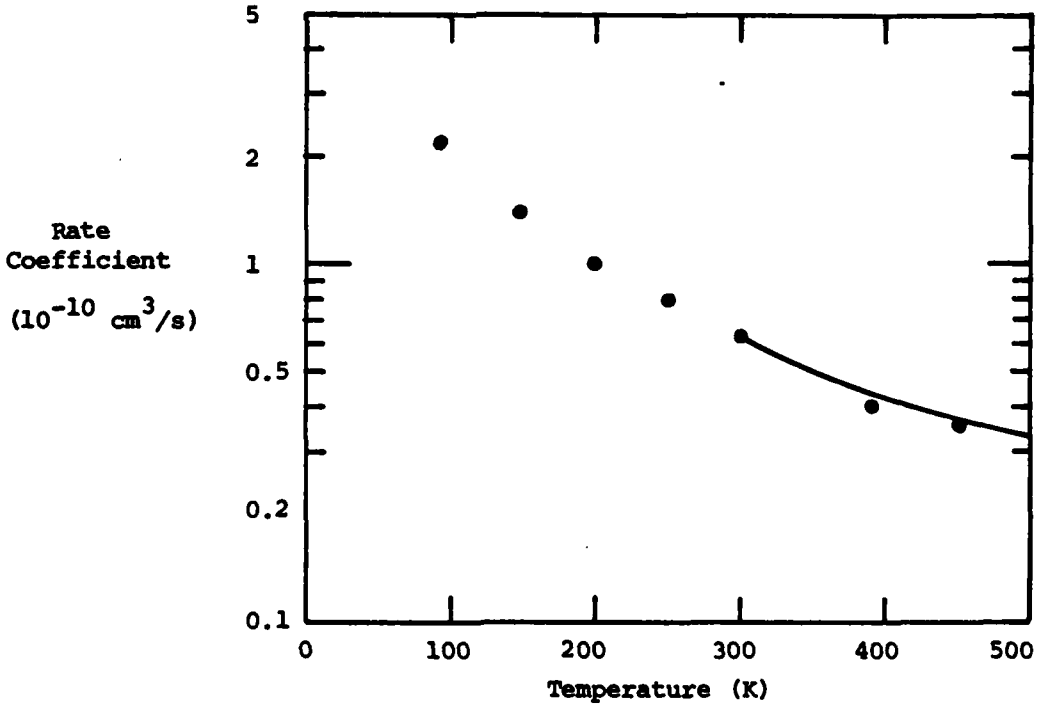


FIG. 4. REACTION RATE COEFFICIENT FOR $CO_2^+ + O_2 \rightarrow CO_2 + O_2^+$. The solid line is an average of various data reported by Lindinger et al.[5]

The rate coefficient for the charge transfer reaction $CO^+ + O_2 \rightarrow CO + O_2^+$ was found to increase with decreasing temperature, approaching the calculated gas kinetic rate of 7.7×10^{-10} cm^3/s at very low temperatures. This reaction is exoergic by 1.950 eV.

The rate coefficient for the charge transfer reaction $CO_2^+ + O_2 \rightarrow CO_2 + O_2^+$ was likewise found to increase with decreasing temperature, approaching the calculated gas kinetic rate of 6.9×10^{-10} cm^3/s at very low temperatures. Our data join smoothly onto the high temperature data of Lindinger et al.,[5] which extend to 900 K and show that the rate coefficient passes through a minimum in the neighborhood of 680 K, implying that a second reaction mechanism becomes effective at higher temperatures. The charge transfer reaction is exoergic by 1.706 eV. All rearrangement reaction channels are forbidden energetically.

The reaction rate coefficients presented above are estimated to be accurate to $\pm 30\%$. The reaction temperatures are estimated to be correct within 5 K, which is the greatest variation observed among five platinum resistance thermometers mounted along the flow tube. However, at high temperatures this variation was as large as 20 K and hence the temperatures above room temperature are much less certain.

B. Ion Flow Velocities and Injection Shock Effects. As a part of our studies of the behavior of the flowing ion swarms at different temperatures we measured the time-of-flight of the ion swarms. These data are of interest to others using SIFT or flowing afterglow apparatuses for two reasons. First, the ion velocity is needed in order to determine the reaction rate coefficient in Eq. (3), and we are not aware of any work done on this matter other than at room temperature.

Second, the results give evidence for undesirable effects of a shock wave in the flow tube associated with the supersonic helium injector. The SIFT injection shock was brought to general attention by Dupeyrat et al.[6] who studied various injector designs in a flow tube and a wind tunnel. The shock apparently carries ions downstream in the flow tube much faster than they would otherwise travel. The shock presents no problem provided (a) the flow has reached equilibrium by the time the ions arrive at the reactant gas titrant port, and (b) the ion velocity is measured entirely in the equilibrium flow region.

Dupeyrat et al.[6] find that the persistence of the shock downstream in the flow tube depends on the ratio of the gas pressures on either side of the injector. A large pressure drop across the injector means that a greater distance must be allowed for the flow to reach equilibrium. (In the AFGL SIFT apparatus the first titrant port is 34 cm downstream of the helium injector.) We have measured the average ion velocity over the first half of the flow tube and over the second half of the flow tube under widely varying conditions. Some of these data are presented in Table I and have been used by us as a diagnostic to indicate safe conditions for ion-molecule reaction studies.

Also calculated for Table I is the ratio of the average equilibrium ion velocity v_i to the bulk helium velocity v_{He} (determined from the helium throughput and pressure). The ion swarm could be pulsed for time-of-flight data either in the ion source or at an electrode midway down the flow tube. We found that the measured time-of-flight is independent of the precise pulsing scheme (turning the ion current on or off with the pulse, partly or fully) and independent of the mass of the ions (from 12 to 44 amu), within about 1%. The less massive ions

TABLE I. A SAMPLE OF ION VELOCITY RESULTS VERSUS FLOW TUBE PRESSURE AND TEMPERATURE.

Temperature (K)	Pressure (torr)	$v_i^{(1)}$ for first half of flow tube	$v_i^{(1)}$ for second half of flow tube	$\frac{v_i^{(1)}}{v_i^{(2)}}$	$\frac{v_i^{(2)}}{v_{He}}$
92	0.160	5864 cm/s	4909 cm/s	1.19	2.18
	0.200	7687	5385	1.43	2.08
	0.230	8402	5579	1.51	2.10
	0.299	13677	6015	2.27	2.10
	0.372	25293	6608	3.83	2.16
250	0.240	15263	13230	1.15	1.72
	0.320	18689	14040	1.33	1.71
	0.372	21046	14637	1.44	1.71
450	0.460	28649	25141	1.14	1.64
	0.550	32971	26296	1.25	1.68
	0.600	34266	26705	1.28	1.68

clearly diffused more in the pulse, however, as evidenced by the greater width of the arrival time spectra.

Considering only the results for v_i from the equilibrated flow, we find (a) the ratio v_i/v_{He} is independent of pressure at a given temperature, and (b) the ratio v_i/v_{He} is constant for temperatures

above about 270 K but increases at decreasing temperatures below 270 K. The ratio v_i/v_{He} is expected to be greater than unity since the ion density is greatest along the axis of the flow tube where the helium velocity is the greatest.[1,7] However, it has heretofore been assumed that the parabolic velocity profile across the diameter of the flow tube was essentially the same at all temperatures, and therefore the ratio v_i/v_{He} would be independent of temperature. Our data indicate that the helium velocity profile across the flow tube becomes more sharply peaked at low temperatures, leading to a larger value of v_i and hence of v_i/v_{He}. This conclusion is supported by viscous flow measurements we have made which imply that a molecular slip contribution is more important at low temperatures. Sufficient molecular slip would decrease the radius of the region of viscous flow and modify the helium velocity profile across the tube.[8] Further analysis of the flow problem is needed at this point before we can offer a quantitative explanation of our results.

C. Other Ion-Molecule Reaction Data. We have obtained a few preliminary results for N^+ and N_2^+ ions reacting with O_2, CO, and NO and for C^+ + NO. These studies are in too early a stage to report on here. Data for C^+ + NO and N^+ + CO are needed for comparison to recent atomic beam results.[9]

IV. RECOMMENDATIONS:

An abstract for a paper on this research has been submitted for presentation at the Gaseous Electronics Conference in Dallas in

October 1982,[10] and we anticipate that one journal article will be prepared on the ion-molecule reaction results and a second article will be written on the implications of the SIFT ion velocity data we have obtained.

We still need to carry out some analytic work in regard to the helium flow problem, as discussed in Sect. III-B, with the goal of supporting our ion velocity data with calculations. These results are important for research with flow-tube reactors (SIFT and flowing afterglow). Very little work has been done at low temperatures, but as the number of flow-tube reactors in use around the world grows, the flow problem must be worked out.

ACKNOWLEDGMENTS

We wish to thank our colleagues at AFGL, John Paulson, Fred Dale, and Tim Thomas, for making our summer exciting and scientifically rewarding; for trusting us with the SIFT apparatus; and for their help minimizing the problems of settling into a new environment. We also thank the Air Force Systems Command, the Air Force Office of Scientific Research, and the Air Force Geophysics Laboratory for making our participation in this research possible.

REFERENCES

1. N. G. Adams and D. Smith, "The Selected Ion Flow Tube (SIFT): A Technique for Studying Ion-Neutral Reactions," Int. J. Mass Spectrom. Ion Phys., Vol. 21, pp. 349-359 (1976).

2. N. G. Adams and D. Smith, "Product-Ion Distributions for Some Ion-Molecule Reactions," J. Phys. B, Vol. 9, pp. 1439-1451 (1976).

3. A. B. Rakshit, H. M. P. Stock, D. P. Wareing, and N. D. Twiddy, "Ion-Molecule Reaction Rates at 100 and 300 K," J. Phys. B, Vol. 11, pp. 4237-4247 (1978).

4. N. G. Adams, D. Smith, and D. Grief, "Reactions of H_nCO^+ Ions with Molecules at 300 K," Int. J. Mass Spectrom. Ion Phys., Vol. 26, pp. 405-415 (1978).

5. W. Lindinger, F. C. Fehsenfeld, A. L. Schmeltekopf, and E. E. Ferguson, "Temperature Dependence of Some Ionospheric Ion-Neutral Reactions from 300-900 K," J. Geophys. Res., Vol. 79, pp. 4753-4756 (1974).

6. G. Dupeyrat, B. R. Rowe, D. W. Fahey, and D. L. Albritton, "Diagnostic Studies of Venturi-Inlets for Flow Reactors," Int. J. Mass Spectrom. Ion Phys., in press.

7. E. E. Ferguson, F. C. Fehsenfeld, and A. L. Schmeltekopf, "Flowing Afterglow Measurements of Ion-Neutral Reactions," Adv. At. Mol. Phys., Vol. 5, pp. 1-56 (1969).

8. E. H. Kennard, Kinetic Theory of Gases (New York, McGraw-Hill,

1938), p. 294.

9. W. Frobin, Ch. Schlier, K. Strein, and E. Teloy, "Ion–Molecule Reactions of N^+ with CO: Integral Reactive Cross Sections in the Collision Energy Range 0.2–13 eV," J. Chem. Phys., Vol. 67, pp. 5505–5516 (1978); also, private communication.

10. T. M. Miller, R. E. Wetterskog, and J. F. Paulson, "Reaction Rate Coefficients for C^+, CO^+, and $CO_2^+ + O_2$," Bull. Am. Phys. Soc., to be published.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

A GaAs FET Computer Model With Applications to Magnetic Field Effects

Prepared by: Michael J. Moloney

Academic Rank: Professor of Physics

Department and University: Rose-Hulman Institute of Technology, Terre Haute, Indiana

Research Location: Electronics Research (Dr. Philip Stover), Avionics Laboratory (AFWAL/AADR), Wright-Patterson AFB, Ohio.

USAF Research Colleague: Dr. David C. Look

Date: 17 August 1982

Contract No: F49620-82-C-0035

A GaAs FET Computer Model With Applications to Magnetic Field Effects

by

Michael J. Moloney

ABSTRACT

A computer model of a GaAs FET has been constructed. It uses explicit charges on the electrodes, and real-time development of the charge distribution to steady state. It shows interesting and realistic I-V relations. Its magnetoresistance results are preliminary but mimic literature results in part. A somewhat novel magnetic field measurement for FET's is proposed. If this effect is large enough to use, the computer model should be very helpful in its interpretation.

I. INTRODUCTION:

I am a college physics professor who has taught an Electronics of Solids course for senior EEs and physics majors for the last four years. This has led me to a strong interest in understanding device physics and made it natural to associate with Dr. Philip Stover's group in the Avionics Laboratory at Wright-Patterson AFB.

This group characterizes GaAs quality by methods including infrared fourier-transform spectroscopy, laser-optical techniques, particle backscatter, C-V, and galvanomagnetic evaluation of transport properties. The group grows much of its material, and is moving toward device applications in heterostructures and FET structures.

This is a rich environment in which to live for ten weeks.

II. OBJECTIVES OF THE RESEARCH EFFORT:

One objective was to study the papers of Fukui[1] and of Lehovec and Zuleeg[2] to see how one might use curve tracer measurements to determine channel mobility, series resistance, etc., of a GaAs FET.

Another objective was to study the work of Wallis and Jay[3] on FET magnetoresistance and magnetotransconductance. This work showed a drop in electron mobility as the channel became pinched off. Measurements like theirs on FET structures were planned.

The overall context of the summer effort was to obtain a better understanding of GaAs FET physics.

III. EVOLUTION OF SUMMER ACTIVITY IN RELATION TO OBJECTIVES:

The experimental work, both on curve tracer and magnetoresistance measurements, ran into early difficulties. A fair amount of preliminary curve tracer work had been carried out when it was firmly determined that the curve tracer malfunctioned under some conditions, and needed repair. That repair, scheduled to be completed in two weeks, stretched out through the summer. Thus, after late June, curve tracer work could be done only with extreme inconvenience.

The early magnetoresistance measurements were made on packages which turned out to be magnetic, and so were thrown about when a magnetic field was applied. This was discovered in late June, and it was not until the first week in August that non-magnetic samples were fabricated. These are just now becoming ready to test.

On the theoretical side, the Fukui[1] paper was based in part on earlier work by Grebene and Ghandi[4], so this earlier paper was studied in some detail. It was a plausible enough analysis, but based on quite unproven assumptions about channel width, and it did not develop any physical feeling in this reader for FET operation.

More literature searching was done[10] and turned up two computer models[5,6] done on FETs shortly after the paper of Grebene and Ghandi. This, plus considerable thought about the whole problem of FET analysis, suggested a model not at all like the others, which used relaxation methods and charge-neutral boundaries.

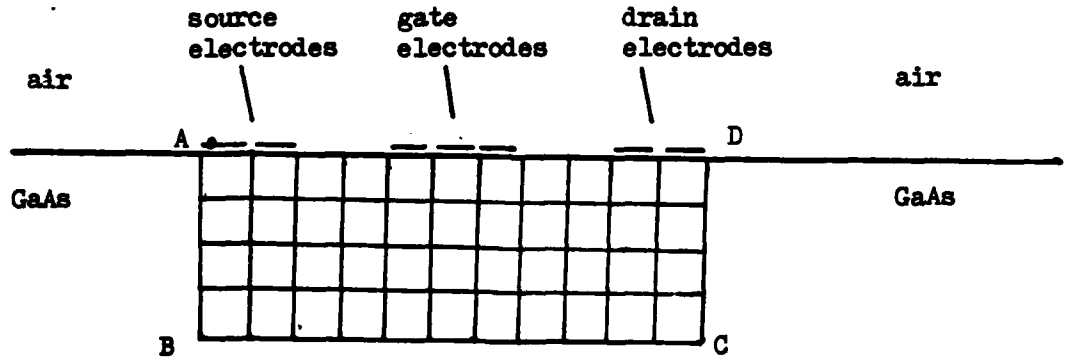
The new model would explicitly include electrodes, and adjust the electrode charges to keep electrode voltages at their specified values. All charges, in the semiconductor and on the electrodes, would be known at each instant of time in this model, and the fields from the charges could be readily calculated and used to help determine the flow of charge.

Because of the electrode charges, this model could generate capacitance information which the others could not. It would be built so that galvanomagnetic effects like magnetoresistance could be a natural part of the model - whose chief role might be the interpretation of these effects.

Essentially all of July 1982 was required to get the model to the edge of realistic functioning. At this point, one had reasonable results for constant electron mobility. Another week was required to successfully include simple velocity saturation. Since then, the velocity overshoot[7,8,9] truly characteristic of GaAs, and scaling changes have been worked on but not completed.

IV. DISCUSSION OF THE MODEL:

The general configuration of this two-dimensional model is shown below. The doped device structure of the model lies in region ABCD of the semi-insulating GaAs half-plane. The model consists of very long cells of square cross-sectional area. Electrodes are located as shown along the interface, modeled by thin sheets of uniform charge/area which will adjust to satisfy the prescribed electrode voltages. There is acceptor doping and electron depletion immediately under the gate, and donor doping in the rest of region ABCD. Each cell has variables of electron concentration and electric potential, and no charge is permitted to flow across the boundary ABCD except at the electrodes.



Each time step begins with all charges known in the GaAs within region ABCD. Initially, this means depletion immediately under the gate, and neutral elsewhere.

The known charges in the semiconductor are used to calculate the electric potential at each electrode position. Then electrode charges are figured out (via some matrix operations) to satisfy the given voltages on source, gate, and drain. At this point, all charges are known, and all boundary conditions are satisfied, so electric fields and potentials can be calculated everywhere.

Care is needed, however, due to the semi-infinite air-GaAs interface[11]. Potentials in the semiconductor due to the electrode sheet charges may be obtained correctly from a single reduced value of the charge, but electric potentials in the semiconductor due to charges within the GaAs require an image charge on the other side of the interface, in addition to a reduced value of the semiconductor charge itself.

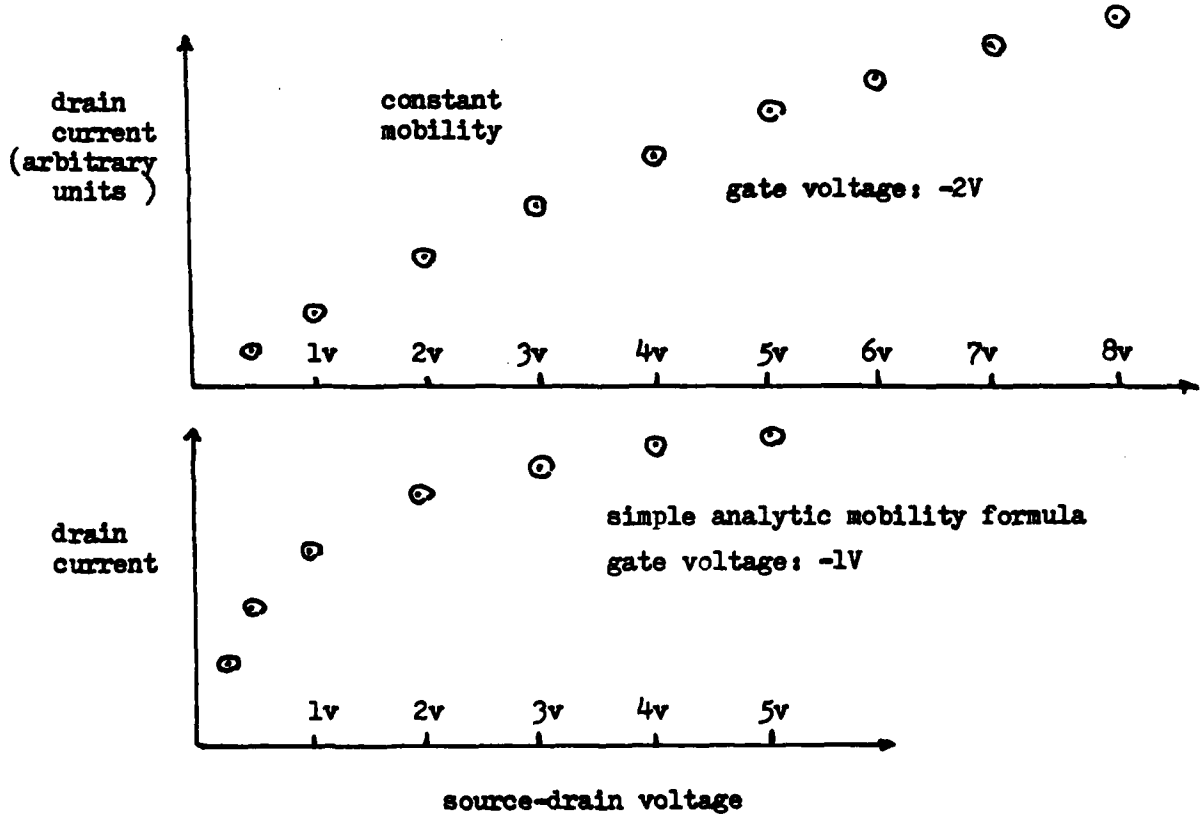
When the electric potential is correct in every cell, the flow of charge in a small time across the four (or fewer) faces of each cell is computed due to drift and diffusion. Cell charges are modified due to the flow (taking care to conserve charge except at electrodes), and the next time step is ready to begin.

The total current, flowing to the right of each vertical column in the model is automatically computed. When column-to-column values of current are close together, steady state is being approached.

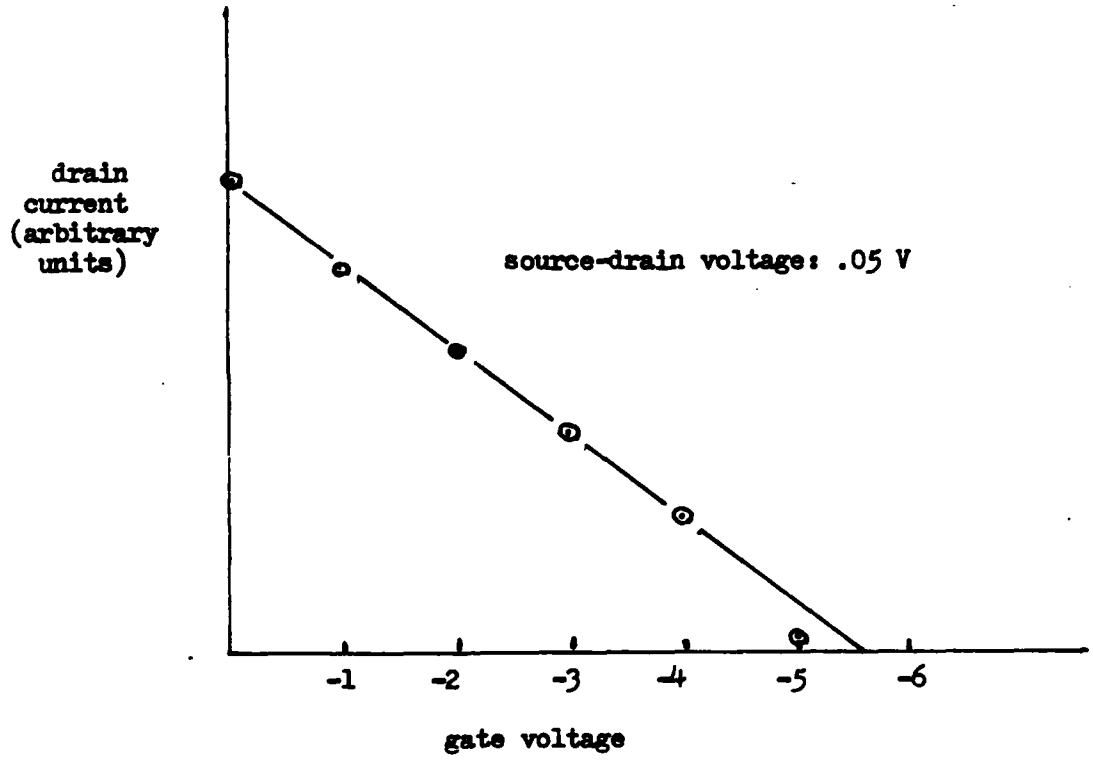
V. MODEL RESULTS:

In a version 20 cells long and 5 cells deep, the model gave a number of satisfactory results. Cell length was 0.25 microns, acceptor doping under the gate was 2×10^{16} cm^{-3}, and donor doping was 4×10^{15} cm^{-3}. These values were chosen for convenience in building and testing the model, rather than for realistic device parameters.

Curves of drain current vs. source-drain voltage at fixed gate voltage were obtained for two cases. The first case, constant mobility no matter what the electric field, is quite simple though not physically realistic, and was used initially to get the model running. The second case was a simple analytic mobility formula used by Lehovec and Zuleeg[2] to model velocity saturation in GaAs. These curves are sketched below.



A series of runs at low source-drain voltage was done varying the gate voltage. The results, shown below, resemble those of real FETs both in the linear relation away from pinchoff (indicated to be about -5.5 V on the graph) and the departure from linearity as pinchoff is approached.



Magnetoresistance was put in the model following a very simple line of reasoning[12]. A particle travelling in crossed electric and magnetic fields travels roughly at a small angle to the electric field given by

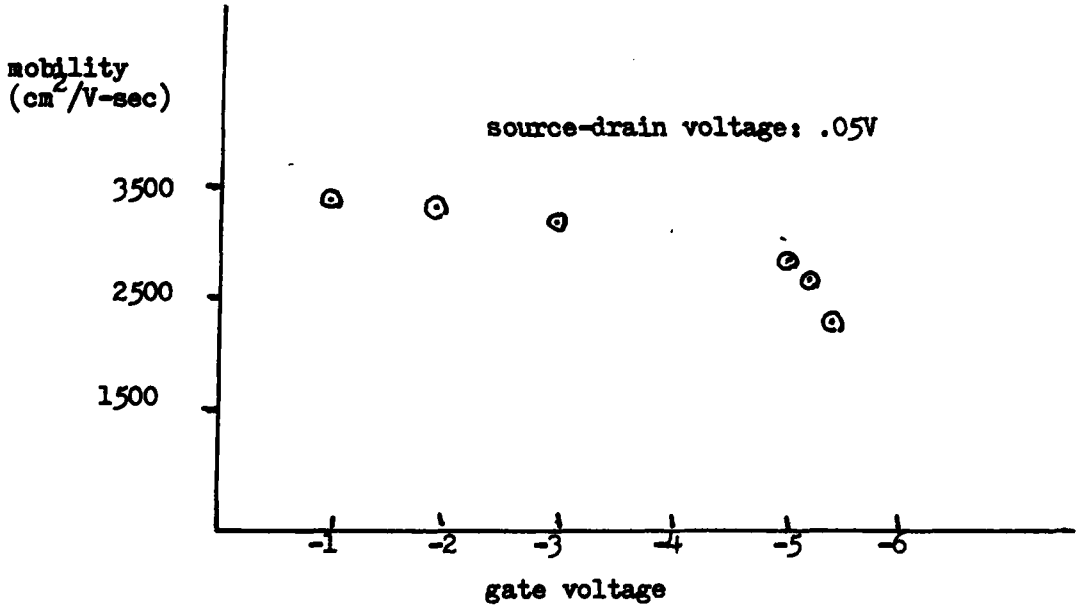
$$\frac{vB}{E} = \mu B \quad ,$$

where v is the drift velocity, B the magnetic field, E the electric field, and μ the mobility.

The fraction of the drift velocity along the electric field direction, turns out, after some averaging, to be

$$1 - \mu^2 B^2 \quad .$$

This correction factor gave model results for low source-drain voltage and gate voltage near pinchoff quite similar to those seen by Wallis and Jay[3]: a decrease of mobility as pinchoff is approached. A sketch of these model results is given below.



VI. RECOMMENDATIONS:

The model has some exceptionally strong points on which to build. With its explicit electrode charges, it makes good physical sense at each instant of time (in contrast to models which make no sense whatever until all iterations are complete). Thus it can give capacitance information and deal with transient and oscillatory situations, though neither of these was needed in the present study. It was built with galvanomagnetic effects in mind, and may have important things to say in this area when it is refined.

The model needs a number of major improvements, which should be readily made as the school year at Rose-Hulman progresses. The true GaAs velocity overshoot[7-9] must be implemented, as well as the ability to model with non-square rectangular cells, in order to get precise depth information about FETs without excessive computation. The time steps to steady state have evolved on a trial-and-error basis. This must be improved, and so must the handling of charge flow through the electrodes.

The experimental GaAs FET magnetoresistance/transconductance measurements still very much need to be done. The computer model has already suggested that the mobility drop seen by Wallis and Jay may well be partly due to high electric fields, even when the source-drain voltage is only 50 mV.

Another possible line of investigation occurred to the author as this report was being written. It has been briefly discussed with Dr. Look, but not literature-searched, so it may or may not be already known.

Microwave FET gates in GaAs are much longer than they are wide. This makes them quite suitable for magnetoresistance work, but of no value for Hall Effect measurements, even if electrodes could be attached in the proper places.

In the conventional Hall Effect arrangement, a magnetic field would be applied perpendicular to the plane containing the electrodes. But if one applied the magnetic field along the long direction of the gate, charge carriers flowing under the gate would be forced either up or down, closer to the gate, or away from it. This amounts to a 'Hall voltage' applied either aiding or opposing the gate voltage. This voltage is easily measured as the difference in gate voltages needed to obtain the same drain current with and without the magnetic field.

The size of this proposed voltage is the product of drift velocity, magnetic field strength, and conducting channel width. Since drift velocity magnitudes are about 10^5 m/s, a channel width of 10^{-8} m and a magnetic field strength of 1 tesla (10 kG) would give a millivolt result. If this could be reliably measured, it might give extremely useful information about channel width near pinchoff.

ACKNOWLEDGEMENTS

This summer's effort in the relaxed yet stimulating atmosphere of AFWAL/AADR has been very beneficial to the author. He learned a great deal, did some things, and developed a lot of ideas.

It is a pleasure to acknowledge the support of the Air Force Systems Command, Air Force Office of Scientific Research, SCEEE, and the people of Dr. Philip Stover's team, especially Dr. David Look.

REFERENCES

Conferences and journal publications:

1. H. Fukui, Bell Syst. Tech. J., 58, 771 (1979).
2. K. Lehovec and R. Zuleeg, Solid State Electron., 13, 1415 (1970).
3. R. H. Wallis and P. R. Jay, Inst. Phys. Conf. Ser. No. 63: Ch. 9 Int. Symp. GaAs and Related Compounds, Japan, 1981.
4. A. B. Grebene and S. K. Ghandi, Solid State Electron., 12, 573 (1969).
5. D. P. Kennedy and R. R. O'Brien, IBM J. Res. Dev., 14, 95 (1970).
6. B. Himsworth, Solid State Electron., 15, 1353 (1972).
7. J. G. Ruch and G. S. Kino, Phys. Rev., 174, 921 (1968).
8. P. A. Houston and A. G. R. Evans, Solid State Electron., 20, 197 (1977).
9. P. Smith, M. Inoue, and J. Frey, Appl. Phys. Lett., 37, 797 (1980).

Textbooks:

10. S. M. Sze, Physics of Semiconductor Devices, Ch. 6, 2nd Ed., Wiley, New York, 1980.
11. J. D. Jackson, Classical Electrodynamics, Sect 4.4, 2nd Ed., Wiley, New York, 1980.
12. P. S. Kireev, Semiconductor Physics, Sect 41.2, Mir Publishers, Moscow, 1975.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

FLUTTER TAMING BY NONLINEAR ACTIVE CONTROL

Prepared by:	Luigi Morino
Academic Rank:	Professor
Department and University:	Dept. of Aerospace and Mechanical Engineering Boston University
Research Location:	Air Force Armament Laboratory, Munitions Division, Compatibility Branch, Structural Dynamics Section
USAF Research Colleague:	Robert W. Bunton
Date:	September 3, 1982
Contract No:	F49620-82-C-0035

FLUTTER TAMING BY NONLINEAR ACTIVE CONTROL

by
Luigi Morino

ABSTRACT

A new concept for the design of aeroservoelastic systems is introduced: flutter taming by nonlinear active control, i.e., use of nonlinear active control to ensure that the behavior of the system beyond the flutter speed is of benign rather than destructive nature. It is shown that flutter taming is always possible for an aeroservoelastic system that can be represented by a system of nonlinear differential equations (with algebraic nonlinearities). This is accomplished by using a very simple nonlinear control law. It is important to emphasize that the active control system for flutter taming is fully nonlinear, and therefore it does not affect the linear behavior (in particular the stability characteristics) of the system. Hence, flutter taming can be used in conjunction with flutter suppression by active control (i.e., use of linear active control to increase the flutter speed).

ACKNOWLEDGEMENTS

This work was supported by the USAF-SCEEE Summer Faculty Research Program. I wish to thank the Air Force Systems Command, the Air Force Office of Scientific Research, and the Southeastern Center for Electrical Engineering Education for the support. I also would like to thank the Air Force Armament Laboratory, in particular Dr. Donald Daniel and Mr. John Macon for the learning opportunity and the excellent support I have received during my tenure at Eglin Air Force Base. In addition, I would like to thank the Structural Dynamics Section of the Aircraft Compatibility Branch, in particular Wiley Robinson, Robert Bunton, Dave Belk, William Miller, Edward Martin, Charles Denegri, Douglas Chojecki, and Marion Barber for their patience, as well as Judith Owens for her promptness in editing and typing this manuscript.

I. INTRODUCTION

The linear description of flutter and the use of linear active control to increase the flutter speed (flutter suppression) are assumed to be known and therefore neither a description of the phenomenon nor a review of the extensive literature in the field are provided here (see, for instance, References 1, 2 and?). Rather, the phenomenon of interest here is the behavior of an aeroservoelastic system beyond the flutter speed, i.e., in the range for which the linear analysis predicts unstable behavior. The nature of the nonlinear flutter was considerably clarified by the research on nonlinear panel flutter which includes the pioneering works of Libresco (Reference ?) and Kobayashi (Reference 5) as well as the extensive work of Dowell (References 6-10). The results of this research activity indicate that the post-flutter behavior can be characterized as follows: the amplitude of the oscillation of the panel beyond the flutter speed grows in time (in agreement with the instability predicted by the linear analysis) but, in contrast to the linear analysis prediction, the amplitude growth tends to a finite value; the limiting value solution is called limit cycle. The above phenomenon is called benign flutter. If the amplitude grows beyond any limit, the phenomenon is called destructive flutter. The mechanism of nonlinear panel flutter was clarified from a mathematical point of view by this author in Reference 11; because of its extensive bearing on the results presented here, the above work and its extensions are extensively reviewed here. In Reference 11, an analytic expression of the transient response was obtained in the form of an asymptotic solution using a singular perturbation method (known as multiple time-scaling method) which had been developed, apparently independently, by Sandri (Reference 12), Frieman (Reference 13), and Nayfeh (Reference 14),

The results indicate that in the neighborhood of the instability boundary, an approximate solution (with an error of the order of magnitude as the fourth power of the limit cycle amplitude) is obtained for the transient response of the panel; the nature of this solution is characterized by one single real number (indicated with in Reference 11). If such number is positive then the nonlinear terms have a stabilizing effect and they counteract the effect of the linear terms (which are destabilizing above the flutter speed) to yield benign flutter (stable limit cycle). On the other hand, if such a number is negative then destructive flutter occurs: it should be emphasized that in this case even for speeds of flight below the flutter speed (where the linear analysis predicts stable behavior) the amplitude may grow beyond any limit provided that the initial amplitude is "sufficiently" high (from a mathematical point of view this behavior is characterized by the presence of an unstable limit cycle, i.e., a limit cycle reached as time goes to minus infinity: this mathematical description of destructive flutter is presented in details later in this paper).

It should be emphasized that the results of Reference 11 are not limited to panel flutter: as shown in Reference 15, the results apply to any dynamic system which can be represented by a system of nonlinear differential equations with nonlinear terms of algebraic nature. In addition, the analysis can be extended to include the effect of fifth order nonlinearities (Reference 16). Finally, the same results are obtained using a completely independent technique, the Lie transform method (Reference 17).

It should be noted that most of the results presented here, particularly the presence of a steady-state (limit-cycle) solution, can be obtained without the use of the multiple-time-scaling method (which, on the other hand, is needed to study whether or not the steady-state solution is stable). The mathematical complexity of the multiple-time-scaling method obscures to a certain extent the simplicity of the fundamental idea in the method. Therefore, a simplified derivation is used here to obtain the steady-state solution. The results on the transient results (necessary to discuss the stability issue) are stated in the main body of this report (the proof is enclosed as Appendix).

For the sake of clarity, two different presentations of flutter taming are given here. The first one, presented in Section 5 (after a review of linear and nonlinear flutter analysis for an aeroelastic system without active control, given in Sections 2, 3 and 4), deals with dynamic systems which can be represented by a system of second order differential equations. This formulation, presented in structural dynamics terms , is convenient to study flutter taming by nonlinear active control of an aircraft that is not subject to linear active control (in particular, without flutter suppression): such a system will be referred to as an aeroelastic system, in contrast to an aeroservoelastic system, which includes active control. Since the equations governing active control are usually given as a set of first order differential equations (state-variable approach), in order to study flutter and/or flutter taming of an aircraft subject to active control, it is convenient to recast the mathematical model as a set of first order differential equations. This system is analyzed in Section 6 and 7 dealing with flutter and flutter taming respectively. The interrelationship between flutter taming and flutter suppression is also discussed in Section 7; in particular, it is shown that

flutter taming does not affect flutter suppression, so that the two concepts are synergistic rather than mutually exclusive. Concluding remarks are given in Section 8.

II. OBJECTIVE

The objectives of the work presented in this report are:

(1) To develop a simple analysis of nonlinear aeroelastic systems that is capable of explaining and predicting the nonlinear behavior of the system, including the type of flutter (i.e., benign or explosive) and the amplitude of the limit cycle.

(2) To introduce a nonlinear control law that ensures that flutter is always benign, and that the steady state amplitude of vibration can be made as small as desired, for instance, smaller than that caused by turbulence (flutter taming).

(3) To study the relationship between linear active control (flutter suppression) and nonlinear active control (flutter taming) and to demonstrate that the two do not interfere with each other and therefore can be used simultaneously.

These items are presented in the body of this report using an intuitive approach; a rigorous derivation of the solution is presented in Appendix A using the two-time scale method of Cole and Kevorkian, reference 18.

III. LINEAR FLUTTER

Consider an aeroelastic system that can be described by a set of equations of the type

$$\ddot{\underline{q}} + (\underline{R} + \lambda \underline{E}_1)\dot{\underline{q}} + (\underline{K} + \lambda \underline{E}_0)\underline{q} = \underline{f}(\underline{q}) \tag{1}$$

or

$$L_\lambda \underline{q} = \underline{f}(\underline{q}) \tag{2}$$

where

$$L_\lambda = \underline{I}\frac{d}{dt^2} + (\underline{R} + \lambda \underline{E}_1)\frac{d}{dt} + \underline{K} + \lambda \underline{E}_0 \tag{3}$$

whereas $\underline{f}$ is the vector of the nonlinear term, which is assumed to be of the type

$$\underline{f}(\underline{q}) = \{f_h\} = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} b_{hij}\, q_i\, q_j + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} c_{hijk}\, q_i\, q_j\, q_k + \cdots \right\} \tag{4}$$

(An outline of the restrictive assumptions implied in the use of Equations 1 and 2 is given in Appendix A.) For the sake of clarity and simplicity assume that the second-order nonlinear terms are equal to zero

$$b_{ijk} = 0 \tag{5}$$

(the case with nonzero second-order terms is just more cumbersome to present than that with zero second-order terms, and does not add any new relevant feature to the analysis, see References 11 and 15).

Next, assume that the amplitude of the solution is very small, say of order ε (where ε is a small parameter to be defined more precisely later as a function of λ)

$$\underline{q} = \varepsilon \underline{q}_1 + O(\varepsilon^2) \tag{6}$$

Combining Equations 1 through 6 yields $\underline{f} = O(\varepsilon^3)$ and

$$L_\lambda \underline{q}_1 = \ddot{\underline{q}}_1 + (\underline{R} + \lambda \underline{E}_1)\dot{\underline{q}}_1 + (\underline{k} + \lambda \underline{E}_0)\underline{q}_1 = 0 \qquad (7)$$

The solution of Equation 7 is a linear combination of elementary solutions of the type

$$\underline{q}_i(t) = \tilde{\underline{q}}_i \, e^{s_i t} \qquad (8)$$

where $\tilde{\underline{q}}_i$ is the nontrivial solution of

$$\underline{M}(s_i, \lambda)\, \tilde{\underline{q}}_i = 0 \qquad (9)$$

with

$$\underline{M}(s, \lambda) = s^2 \underline{I} + s(\underline{R} + \lambda \underline{E}_1) + \underline{k} + \lambda \underline{E}_0 \qquad (10)$$

whereas s_i is the ith root of the characteristic equation

$$Det\left[\underline{M}(s, \lambda)\right] = 0 \qquad (11)$$

(It may be worth noting that $\underline{M}(s, \lambda)$ is the frequency-domain transformed of L_λ.)

For $\lambda = 0$ all the roots are stable (i.e., lie on the left hand side of the complex s-plane). As λ increases one of the roots (and its complex conjugate, if any) eventually crosses the imaginary axis: at the stability (flutter) boundary, $\lambda = \lambda_F$, Real s=0 and

$$s = i\omega_F \qquad (12)$$

Beyond the stability boundary, the system becomes unstable. If the root crosses the imaginary axis at the origine, then at the stability boundary, s=0 and the system does not oscillate; one refers to this instability as divergence. On the other hand, if the root does not cross the imaginary axis at the origine, then at the stability boundary the system oscillates and one refers to this instability as flutter: ω_F is called the flutter frequency, the velocity V_F corresponding to $\lambda_F = \frac{1}{2}\rho_\infty V_F^2$ is called flutter speed. (It may be worth noting that divergence can be thought of as a particular case of flutter, a zero-frequency flutter, if you will.)

The flutter (steady state) solution (note that as time goes infinity all the other elementary solutions go to zero) is given by Equation 4 with

$$\underline{q}_1 = a\,\underline{\tilde{q}}_F\, e^{i\omega_F t} + a^*\,\underline{\tilde{q}}_F^*\, e^{-i\omega_F t} \qquad (13)$$

where $\underline{\tilde{q}}_F$ is the nontrivial solution of Equation 9 with $\lambda = \lambda_F$ and $s = i\omega_F$:

$$\underline{M}\,(i\omega_F, \lambda_F)\,\underline{\tilde{q}}_F = 0 \qquad (14)$$

Note that the flutter mode is a linear combination of the modes of vibrations with coefficients equal to the components of $\underline{\tilde{q}}_F$. The vector $\underline{\tilde{q}}_F$ will be referred to as the flutter vector. In addition, a is an arbitrary complex constant, which depends upon the initial conditions. (As it will be shown in the next Section, this is true only for linear systems. For nonlinear system, the amplitude a (limit-cycle amplitude) is obtained as a part of the solution, independently, in general, of the initial conditions.)

IV. NONLINEAR FLUTTER

In order to study the nonlinear behavior of system under consideration, I will consider a perturbation of the linear flutter solution. Note that $\underline{f}$ is of order ε^3. Since we are looking for a balance between the increase in linear terms (due to a change in λ) and the nonlinear terms, it is necessary that these terms be of the same order of magnitude: this is accomplished by setting

$$\lambda = \lambda_F + \varepsilon^2 \lambda_2 \tag{15}$$

Actually, since ε has been defined except for an arbitrary multiplicative constant I will assume

$$\lambda_2 = \pm 1 \tag{16}$$

so that

$$\varepsilon = \left| \lambda - \lambda_F \right|^{1/2} \tag{17}$$

(Equation 17 clearly indicates that the "balance between increase-in-linear-terms and nonlinear-terms" implies that the amplitude a of the solution grows like the square root of $|\lambda - \lambda_F|$.)

As mentioned above in order to study the nonlinear behavior of the system I will consider a solution which is a perturbation of the linear solution at the flutter boundary. More precisely, knowing that the "perturbing factors" (i.e., nonlinear terms and increase in linear terms) are of order ε^3, I will be looking for a solution of the type

$$\underline{q} = \varepsilon \underline{q}_1 + \varepsilon^3 \underline{q}_3 + \cdots \tag{18}$$

with $\underline{q}_1$ given by Equation 13, where now the term a is not necessarily a constant, but reduces to a constant when ε goes to zero. The reason for this is that as we know, the frequency of vibration changes with λ : in order to allow sufficient flexibility in the solution it is necessary to assume that a

is slowly varying with time; since the objective of this Section is to obtain a steady-state solution, it will suffice to assume that

$$a = a_o \, e^{i\omega_2 \varepsilon^2 t} \tag{19}$$

where a is a complex constant whereas ω_2 is a real constant.

[An arbitrary dependence of a upon a slow time scale is needed to study the actual transient response of the solution (this is necessary for the analysis of the stability of the limit-cycle). It should be emphasized that Equation 15, 18 and 19, introduced here on the basis of an intuitive argument, can actually be obtained by letting the order of the corrections be initially arbitrary and then let the mathematics yield that the correct expressions are those given by Equations 15, 18 and 19: this is accomplished by using, for instance, the multiple-time-scaling method (References 11, 14 and 15) or the Lie transform method (Reference 16). For the analysis presented here (limited to third order nonlinear terms) the two-time-scale method is sufficient to obtain correct results in a mathematically convincing manner (that is, without resorting to physically intuitive arguments): such analysis is presented in Appendix B.]

Combining Equation 18 with Equations 13 and 19 yields

$$\underline{q} = \varepsilon\left(a_o \, \underline{\tilde{q}}_F \, e^{i\omega t} + a_o^* \, \underline{\tilde{q}}_F^* \, e^{-i\omega t}\right) + \varepsilon^3 \underline{q}_3 + \cdots \tag{20}$$

with

$$\omega = \omega_F + \omega_2 \varepsilon^2 + \cdots \tag{21}$$

Substituting Equation 21 into Equation 1 yields

$$\begin{aligned}
&\varepsilon \left[\underline{M}(i\omega, \lambda_F)\, \tilde{\underline{q}}_F\, a_o\, e^{i\omega t} + \right. \\
&\quad \underline{M}(-i\omega, \lambda_F)\, \tilde{\underline{q}}_F^*\, a_o^*\, e^{-i\omega t} + \\
&\varepsilon^3 \left[\lambda_2 (i\omega \underline{E}_1 + \underline{E}_0)\, \tilde{\underline{q}}_F\, a_o\, e^{i\omega t} + \right. \\
&\quad \lambda_2 (-i\omega \underline{E}_1 + \underline{E}_0)\, \tilde{\underline{q}}_F^*\, a_o^*\, e^{-i\omega t} + \\
&\quad \left. \left. \underline{f}(\underline{q}_1) + L_F\, \underline{q}_3 \right] + \cdots \right.
\end{aligned} \tag{22}$$

where L_F is equal to L_λ with $\lambda = \lambda_F$, whereas

$$\begin{aligned}
\underline{f}(\underline{q}_1) = a^3\, \tilde{\underline{f}}_3\, e^{i3\omega t} + a^2 a^*\, \tilde{\underline{f}}_1\, e^{i\omega t} \\
+ a^{*2} a\, \tilde{\underline{f}}_1^*\, e^{-i\omega t} + a^{*3}\, \tilde{\underline{f}}_3^*\, e^{-i3\omega t}
\end{aligned} \tag{23}$$

with

$$\begin{aligned}
\tilde{\underline{f}}_3 &= \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} c_{hijk}\, \tilde{q}_{F_i}\, \tilde{q}_{F_j}\, \tilde{q}_{F_k} \right\} \\
\tilde{\underline{f}}_1 &= \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} c_{hijk} \left(\tilde{q}_{F_i}\, \tilde{q}_{F_j}\, \tilde{q}_{F_k}^* + \right. \right. \\
&\quad \left. \left. \tilde{q}_{F_i}\, \tilde{q}_{F_j}^*\, \tilde{q}_{F_k} + \tilde{q}_{F_i}^*\, \tilde{q}_{F_j}\, \tilde{q}_{F_k} \right) \right\}
\end{aligned} \tag{24}$$

Noting that ω is given by Equation 19 and separating terms of the same order of magnitude yields

$$\begin{aligned}
&\underline{M}(i\omega_F, \lambda_F)\, \tilde{\underline{q}}_F\, a\, e^{i\omega_F t} + \\
&\underline{M}(-i\omega_F, \lambda_F)\, \tilde{\underline{q}}_F^*\, a^*\, e^{-i\omega_F t} = 0
\end{aligned} \tag{25}$$

and

$$L_F(\underline{q}_3) = -a^3 \underline{\tilde{f}}_3 e^{i3\omega_F t} +$$

$$\{[2\omega_F \omega_2 \underline{I} - i\omega_2(\underline{R} + \lambda_F \underline{E}_1) - \lambda_2(i\omega_F \underline{E}_1 + \underline{E}_0)]a\, \underline{\tilde{q}}_F$$

$$- a^2 a^* \underline{\tilde{f}}_1 \} e^{i\omega_F t} + \text{conjugate terms} \tag{26}$$

Equation 25 is automatically satisfied by definition of $\underline{q}_F$ (see Equation 14). On the other hand, the solution of Equation 26 is given by

$$\underline{q}_3 = \underline{q}_{GH} + \underline{q}_3^{(3)} e^{i3\omega_F t} + \underline{q}_3^{(1)} e^{i\omega_F t}$$

$$+ \text{conjugate terms} \tag{27}$$

where $\underline{q}_{GH}$ is the general solution of the homogeneous equation (identical to Equation 9) whereas $\underline{\tilde{q}}_3^{(3)}$ and $\underline{\tilde{q}}_3^{(1)}$ are the solutions of

$$\underline{M}(i3\omega_F, \lambda_F)\, \underline{\tilde{q}}_3^{(3)} = -a^3 \underline{\tilde{f}}_3 \tag{28}$$

$$\underline{M}(i\omega_F, \lambda_F)\, \underline{\tilde{q}}_3^{(1)} = \underline{\tilde{z}}_3^{(1)} \equiv$$

$$[2\omega_F \omega_2 \underline{I} - i\omega_2(\underline{R} + \lambda_F \underline{E}_1) - \lambda_2(i\omega_F \underline{E}_1 + \underline{E}_0)]a\, \underline{\tilde{q}}_F - a^2 a^* \underline{\tilde{f}}_1 \tag{29}$$

Note that, for $\lambda = \lambda_F$, the only imaginary root of the characteristic equation (Equation 11) is $s = i\omega_F$. Hence, the determinant of Equation 28 is different from zero (and therefore the solution exists and is unique). On the other hand, the determinant of Equation 29 is equal to zero. This implies that there exist a linear combination (with coefficients c_i) of the rows of $\underline{M}(i\omega_F, \lambda_F)$ which is identically equal to zero: in other words, there exist row-matrix $\underline{\tilde{q}}_L^T$ where $\underline{\tilde{q}}_L = \{c_i\}$ such that

$$\underline{\tilde{q}}_L^T\, \underline{M}(i\omega_F, \lambda_F) = 0 \tag{30}$$

The solution to Equation 29 exists if and only if the same linear combination of the right hand side is also equal to zero, i.e.,

$$\underline{\tilde{q}}_L^T\, \underline{\tilde{z}}_3^{(1)} = 0 \tag{31}$$

(The solution is not unique; however, the arbitrary additional term in the solution is already included in $\underline{q}_{GH}$ and therefore need not be considered here).

Equation 31 is the key to the nonlinear analysis, since it is the condition that determines the value of the limit-cycle amplitude. [It may be worth noting that $\tilde{\underline{q}}_L^T$ is the left characteristic vector of $M(s, \lambda_F)$ corresponding to the characteristic value $s = i\omega_F$ (in other words, $\tilde{\underline{q}}_L$ is the characteristic vector of the transpose of $\underline{M}(i\omega_F, \lambda_F)$; compare Equation 30 to Equation 14). Because of an orthogonality property of left and right characteristic vectors, Equation 30 implies that $\underline{z}_3^{(1)}$ has as no component in the direction of the right-characteristic-vector, $\tilde{\underline{q}}_F$, and therefore does not "excite resonance", so to speak.]

In order to show how Equation 31 yields the limit-cycle amplitude, combine Equation 31 with the expression for $\underline{z}_3^{(1)}$, Equation 29, to obtain

$$i\omega_2 \tilde{\underline{q}}_L^T (2i\omega_F \underline{I} + \underline{R} + \lambda_F \underline{E}_1) \tilde{\underline{q}}_F a + \lambda_2 \tilde{\underline{q}}_L^T (i\omega_F \underline{E}_1 + \underline{E}_0) \tilde{\underline{q}}_F a + \tilde{\underline{q}}_L^T \tilde{\underline{f}}_1 a^2 a^* = 0 \qquad (32)$$

or

$$i\omega_2 a_0 + \beta a_0 + \gamma a_0^2 a_0^* = 0 \qquad (33)$$

where

$$\beta = \lambda_2 \frac{\tilde{\underline{q}}_L^T (i\omega_F \underline{E}_1 + \underline{E}_0) \tilde{\underline{q}}_F}{\tilde{\underline{q}}_L^T (2i\omega_F \underline{I} + \underline{R} + \lambda_F \underline{E}_1) \tilde{\underline{q}}_F} \qquad (34)$$

$$\gamma = \frac{\tilde{\underline{q}}_L^T \tilde{\underline{f}}_1}{\tilde{\underline{q}}_L^T (2i\omega_F \underline{I} + \underline{R} + \lambda_F \underline{E}_1) \tilde{\underline{q}}_F} \qquad (35)$$

Letting

$$\beta = \beta_R + i\beta_I = \text{Real}(\beta) + i\,\text{Imag}(\beta) \qquad (36)$$

$$\gamma = \gamma_R + i\gamma_I = Real(\gamma) + i\,Imag(\gamma) \qquad (37)$$

and

$$a_o = |a_o| e^{i\varphi} \qquad (38)$$

and separating real parts in imaginary part in Equation 33 finally yields

$$\beta_R |a_o| + \gamma_R |a_o|^3 = 0 \qquad (39)$$

and

$$\omega_2 |a_o| + \beta_I |a_o| + \gamma_I |a_o|^3 = 0 \qquad (40)$$

By choosing λ_2 (see Equation 34) so that

$$\beta_R / \gamma_R > 0 \qquad (41)$$

one obtains

$$|a_o| = |\beta_R / \gamma_R|^{1/2} \qquad (42)$$

and

$$\omega_2 = -\beta_I - \gamma_I |a_o|^2 \qquad (43)$$

$$= -\beta_I + \gamma_I \beta_R / \gamma_R$$

which give, respectively the amplitude of the limit cycle and the change in frequency of vibration.

V. SUMMARY OF NONLINEAR FLUTTER ANALYSIS

In summary, the solution may be written as (see Equation 20)

$$\underline{q}(t) = |\lambda - \lambda_F|^{1/2} |a_0| 2 \, \text{Real} \left(\tilde{\underline{q}}_F \, e^{i(\omega t + \varphi)} \right) + O(\epsilon^3) \quad (44)$$

where $\tilde{\underline{q}}_F$ is the flutter vector (see Equation 14) whereas (see Equation 41)

$$|a_0| = |\beta_R / \gamma_R|^{1/2} \quad (45)$$

and (see Equation 43)

$$\omega = \omega_F + |\lambda - \lambda_F| \left(\gamma_I \beta_R / \gamma_R - \beta_I \right) + \cdots \quad (46)$$

where β_R, β_I, γ_R and γ_I are the real and imaginary parts of β and γ, which are given by Equations 34 and 35, with $\tilde{\underline{q}}_L$ given by Equation 30 (i.e., the transpose of Equation 14 which defines the flutter vector $\tilde{\underline{q}}_F$) and $\tilde{\underline{f}}_1$ given by Equation 24: all the other matrices are prescribed from the formulation of the problem (see Equation 1). Finally, the value of $\lambda_2 = \pm 1$ (i.e., whether $\lambda > \lambda_F$ or $\lambda < \lambda_F$) is determined by the condition $\beta_R / \gamma_R < 0$ (see Equation 40).

Finally, it may be worth noting that the analysis of the transient response, gives further insight into the solution: these results are all derived from the generalization of Equation 38, which is the case of transient analysis is given by (see Appendix B)

$$\frac{d|a_0|}{d\tau} + \beta_R |a_0| + \gamma_R |a_0|^3 = 0 \quad (\tau = \epsilon^2 t) \quad (47)$$

The first result which is obtained from Equation 47 is that since linear analysis predicts instability for $\lambda > \lambda_F$, then the linear terms are destabilizing for $\lambda > \lambda_F$ (and stabilizing for $\lambda < \lambda_F$): this implies that

$$\beta_R \leq 0 \qquad \text{for } \lambda_2 = \pm 1 \quad (\lambda \gtrless \lambda_F) \quad (48)$$

Hence, Equation 40 is satisfied by choosing λ_2 as follows

$$\lambda_2 = \pm 1 \qquad \text{for} \quad \gamma_R \gtrless 0 \quad (49)$$

In other words, if $\gamma_R > 0$, $\lambda_2 = +1$ and $\lambda = \lambda_F + \varepsilon^2 > \lambda_F$. This implies that a limit-cycle solution exists for $\lambda > \lambda_F$. Equation 47 shows that in this case ($\gamma_R > 0$) the nonlinear terms are stabilizing (since $d|a_0|/d\tau > 0$ without the presence of linear terms) and that $|a_0|$ tends to $|a_0| = |\beta_R / \gamma_R|^{1/2}$ independently of the initial conditions. On the other hand, for $\lambda_2 = -1$ ($\lambda = \lambda_F - \varepsilon^2 < \lambda_F$) both β_R and γ_R are positive and therefore the solution always goes to zero (faster than the linear theory predicts). The parameter is referred to here as the stability parameter of the nonlinear terms.

The other possibility is $\gamma_R < 0$ (for $\gamma_R = 0$ the nonlinear behavior is equal to the linear one, except for the presence of $e^{i3\omega t}$ components): in this case the nonlinear terms are destabilizing and for $\lambda > \lambda_F$ ($\lambda_2 = 1$) both β_R and γ_R are negative and therefore $|a_0|$ goes to infinity (in a finite amount of time, see Appendix B). On the other hand, there exists a limit cycle for $\lambda = \lambda_F - \varepsilon^2 < \lambda_F$: however, the limit cycle solution is unstable; for, as indicated by Equation 47, $|a_0|$ decreases to zero if $|a_0|_0 < |\beta_R / \gamma_R|^{1/2}$, but it grows to infinity if $|a_0|_0 > |\beta_R / \gamma_R|^{1/2}$; in physical terms this indicates that destructive flutter may be experienced even below flutter speed (i.e., for $\lambda < \lambda_F$, where the linear theory predicts stable behavior), provided that the initial conditions are "sufficiently high" (i.e., $|a_0|_0 > |\beta_R / \gamma_R|^{1/2}$).

VI. FLUTTER TAMING OF AN AEROELASTIC SYSTEM

The results of the nonlinear flutter analysis, summarized in Section V, indicate that the nature of flutter (i.e., benign or destructive), beyond flutter speed, is determined by the sign of the real part of γ, as given by Equation 35. This suggests a simple procedure for flutter taming: augment the system so that the additional component in the nonlinear-term vector produces the desired positive value of γ_R. It is also convenient that the linear characteristics of the augmented system be equal to those of the original system. This is necessary in order to avoid, for instance, that the introduction of flutter taming decrease the flutter speed; also, as shown in Section VII, this requirement decouples flutter taming from flutter suppression thereby facilitating the simultaneous implementation of both systems on the same aircraft.

An augmented system that satisfies the above requirements is (see Equation 2 for the original system)

$$L_1 \underline{q} + \underline{b}_0 u + \underline{b}_1 \dot{u} = \underline{f}(\underline{q}) + \hat{\underline{f}}(\underline{q}, u) \qquad (50)$$

$$\ddot{u} + a_1 \dot{u} + a_0 u = \alpha \left(\underline{c}_0^T \underline{q} + \underline{c}_1^T \dot{\underline{q}} \right)^3 \qquad (51)$$

From a physical point of view, this modification can be obtained by letting u be the deflection of a control surface, the dynamics of which is given by the left-hand side of Equation 51 (the control surface is assumed to be irreversible, that is its dynamics is not affected by the aerodynamic forces). This deflection produces generalized forces which are approximated as $\underline{b}_0 u + \underline{b}_1 \dot{u} - \hat{\underline{f}}(\underline{q}, u)$ (where $\hat{\underline{f}}$ contains only cubic nonlinear terms); this is consistent with the approximation of the aerodynamic forces used in Equation 1 (see Equation C.9). Finally, the feedback is given on the right-hand side of Equation 51 and is proportional to the cube of a linear combination of $\underline{q}$ and $\dot{\underline{q}}$ (e.g., the output

of accelometer integrated once or twice).

The system governed by Equations 50 and 51 is desired augmented system to implement flutter taming. Equations 50 and 51 may be recast as

$$L_\lambda^+ \; \underline{q}^+ = \underline{f}^+ (\underline{q}^+) \tag{52}$$

where (using partitioned-matrix notations)

$$\underline{q}^+ = \left\{ \begin{array}{c} \underline{q} \\ \cdots \\ u \end{array} \right\} \tag{53}$$

$$L_\lambda^+ = \left[\begin{array}{c:c} L_\lambda & \lambda \, \underline{b}_0 + \underline{b}_1 \frac{d}{dt} \\ \hdashline 0 & \frac{d^2}{dt^2} + a_1 \frac{d}{dt} + a_0 \end{array} \right] \tag{54}$$

and

$$\underline{f}^+ = \left\{ \begin{array}{c} \underline{f} + \hat{\underline{f}} \\ \cdots \\ \alpha \, (\underline{c}_0^T \underline{q} + \underline{c}_1^T \dot{\underline{q}})^3 \end{array} \right\} \tag{55}$$

Equation 52 is formally equal to Equation 2: therefore the results obtained for Equation 2 are applicable to Equation 52 as well.

In particular, the linear analysis (see Section 2, Equation 11) yields that the solution is obtained by setting

$$\det \left[\underline{M}^+ (s, \lambda) \right] = 0 \tag{56}$$

where $\underline{M}^+(s, \lambda)$ is the frequency-domain transformed of L_λ, i.e.,

$$\underline{M} (s, \lambda) = \left[\begin{array}{c:c} \underline{M}(s, \lambda) & \lambda (\underline{b}_0 + s \underline{b}_1) \\ \hdashline 0 & s^2 + a_1 s + a_0 \end{array} \right] \tag{57}$$

where

$$\equiv s^2 \underline{I} + s (\underline{R}^+ + \lambda \underline{E}_1^+) + \underline{K}^+ + \lambda \underline{E}_0^+$$

$$\underline{R}^+ = \left[\begin{array}{c:c} \underline{R} & \underline{0} \\ \hdashline \underline{0} & a_1 \end{array} \right] \tag{58}$$

$$\underline{K}^{+} = \left[\begin{array}{c|c} \underline{K} & \underline{0} \\ \hline \underline{0} & a_0 \end{array}\right] \tag{59}$$

$$\underline{E}_0^{+} = \left[\begin{array}{c|c} \underline{E}_0 & \underline{b}_0 \\ \hline \underline{0} & 0 \end{array}\right] \tag{60}$$

$$\underline{E}_1^{+} = \left[\begin{array}{c|c} \underline{E}_1 & \underline{b}_1 \\ \hline \underline{0} & 0 \end{array}\right] \tag{61}$$

Therefore

$$\det\left[\underline{M}^{+}(s,\lambda)\right] = (s^2 + a_1 s + a_0)\ \det\left[\underline{M}(s,\lambda)\right] \tag{62}$$

which indicates that the eigenvalues of the augmented system are the same as those of the original system.

In particular for $\lambda=\lambda_F$ and $s=i\omega_F$ (Equation 12) one obtains

$$\underline{M}^{+}\left(i\omega_F, \lambda_F\right) \underline{\tilde{q}}_F^{+} = 0 \tag{63}$$

which yields (see Equation 14)

$$\underline{\tilde{q}}_F^{+} = \left\{\begin{array}{c} \underline{\tilde{q}}_F \\ \hline 0 \end{array}\right\} \tag{64}$$

Therefore, the steady-state solution (see Equation 13)

$$\underline{q}^{+} = a\, \underline{\tilde{q}}_F^{+}\, e^{i\omega_F t} + a^{*}\, \underline{\tilde{q}}_F^{+\,*}\, e^{-i\omega_F t} \tag{65}$$

is equal that of the original system (except for the added component of u which is equal to zero).

On the other hand, in order to obtain the nonlinear solution we need $\underline{\tilde{q}}_L^{+}$ defined by (compare to the transpose of Equation 30)

$$\left[\underline{M}^{+}(i\omega_F, \lambda_F)\right]^{T}\ \underline{\tilde{q}}_L^{+} = 0 \tag{66}$$

or, using the definition of $\underline{M}^{+}$ (see Equation 57)

$$\left[\begin{array}{c|c} \underline{M}^T(i\omega_F, \lambda_F) & \underline{0} \\ \hline \lambda_F(\underline{b}_0^T + i\omega_F \underline{b}_1^T) & -\omega_F^2 + i\omega_F a_1 + a_0 \end{array}\right] \underline{\tilde{q}}_L^+ = 0 \qquad (67)$$

which is satisfied by

$$\underline{\tilde{q}}_L^+ = \left\{\begin{array}{c} \underline{\tilde{q}}_L \\ \hline \sigma \end{array}\right\} \qquad (68)$$

where

$$\sigma = \frac{-1}{-\omega_F^2 + i\omega_F a_1 + a_0} \lambda_F (\underline{b}_0^T + i\omega_F \underline{b}_1^T) \underline{\tilde{q}}_L \qquad (69)$$

which is different from zero if

$$(\underline{b}_0^T + i\omega_F \underline{b}_1^T) \underline{\tilde{q}}_L \neq 0 \quad \text{and} \quad -\omega_F^2 + i\omega_F a_1 + a_0 \neq 0 \qquad (70)$$

(i.e., ω_F is not a natural frequency of the system representing the control-surface dynamics).

The nonlinear solution is given by (see Equation 44)

$$\underline{q}^+(t) = |\lambda - \lambda_F|^{1/2} |a_0|^+ 2\,\mathrm{Real}\left(\underline{\tilde{q}}_F^+ e^{i(\omega^+ t + \varphi)}\right) + \ldots \qquad (71)$$

where

$$|a_0^+| = |\beta_R^+ / \gamma_R^+| \qquad (72)$$

and

$$\omega^+ = \omega_F + |\lambda - \lambda_F| (\gamma_I^+ \beta_R^+ / \gamma_R^+ - \beta_I^+) + \ldots \qquad (73)$$

where β_R^+ and β_I^+ are the real and imaginary parts of

$$\beta^+ = \lambda_2 \frac{\underline{\tilde{q}}_L^{+T} (i\omega_F \underline{E}_1^+ + \underline{E}_0^+) \underline{\tilde{q}}_F^+}{\underline{\tilde{q}}_L^{+T} (2i\omega_F \underline{I} + \underline{R}^+ + \lambda_F \underline{E}_1^+) \underline{\tilde{q}}_F^+} \qquad (74)$$

whereas γ_R^+ and γ_I^+ are the real and imaginary parts of

$$\gamma^+ = \frac{\underline{\tilde{q}}_L^{+T} \underline{\tilde{f}}_1^+}{\underline{\tilde{q}}_L^{+T} (2i\omega_F \underline{I} + \underline{R}^+ + \lambda_F \underline{E}_1^+) \underline{\tilde{q}}_F^+} \qquad (75)$$

Note that the last component of $\underline{\tilde{q}}_F^+$ is equal to zero (Equation 64) and that the last rows of $\underline{I}$, $\underline{R}^T$, $\underline{K}^+$, $\underline{E}_0^+$ and $\underline{E}_1^+$ are equal to zero (except for the last term): this implies

$$\beta^+ = \beta \tag{76}$$

(as it is to be expected because the linear systems are equivalent and β^+ represents the effect of the linear system), whereas

$$\gamma^+ = \gamma + \hat{\gamma} + \alpha \gamma_1 \tag{77}$$

where γ is the effect of the nonlinear terms for the original system (Equation 35), whereas (see Equation 53)

$$\hat{\gamma} = \frac{\tilde{\underline{q}}_L^T \, \hat{\underline{f}}_1}{\tilde{\underline{q}}_L^T \, (2 i \omega_F \underline{I} + \underline{R} + \lambda_F \underline{E}_1) \, \tilde{\underline{q}}_F} \tag{78}$$

is the contribution due to the nonlinear aerodynamic forces due the control surface and (see Equation 55)

$$\gamma_1 = 3 \sigma \left[(\underline{C}_0^T + i \omega_F \underline{C}_1) \, \tilde{\underline{q}}_F \right]^2 \left[(\underline{C}_0^T - i \omega_F \underline{C}_1) \, \tilde{\underline{q}}_F^* \right] \tag{79}$$

Equation 77 is the key to flutter taming. If

$$\gamma_{1_R} \neq 0 \tag{80}$$

and $\sigma \neq 0$ (see Equations 69 and 70), the condition for benign flutter

$$\gamma_R^+ = \gamma_R + \hat{\gamma}_R + \alpha \gamma_{1R} > 0 \tag{81}$$

can always be satisfied by an appropriate value of α.

Note also that α can be chosen so that γ_R^+ is as large as desired: this implies (see Equation 71) that the amplitude of the limit-cycle solution is as small as desired. In other words, by suitable choice of the sign and magnitude of α it is possible to accomplish that:

1. Flutter is benign.

2. Amplitude of vibration is as small as desired.

VII. FLUTTER OF AEROSERVOELASTIC SYSTEMS

An alternative derivation of the result of Section VI is presented here. The advantages of this derivation are that:

1. It is formally simpler and more elegant than that of Section 3.

2. It is applicable when the nonlinear terms are functions of $\underline{q}$ as well as $\dot{\underline{q}}$.

3. It is applicable to systems of equations which are not only of second order (in particular to aeroelastic systems subject to active control, e.g., for flutter suppression).

The advantage of the derivation of Section IV is that it makes it easier (for me, at least) to follow the physical phenomenon thereby facilitating the "creative process", i.e., choosing the control law for flutter taming.

This alternative formulation is obtained by recasting the governing equations as a set of first order differential equations. This yields (see Equation A.20) a system of the type

$$\dot{\underline{x}} = (\underline{A}_0 + \lambda \underline{A}_1)\,\underline{x} + \underline{f}(\underline{x}) \qquad (82)$$

The state-variables x_i include the generalized Lagarangian coordinates, q_i , their time derivatives, $\dot{q}_i$, and all the control variables (if any), introduced for instance for flutter suppression.

Assuming $\underline{f}$ to have only cubic nonlinearities, the solution is of the type

$$\underline{x} = \varepsilon\, \underline{x}_1 + \varepsilon^3 \underline{x}_3 + \cdots \qquad (83)$$

Combining Equation 82 and 83 yields

$$\varepsilon \left[\underline{\dot{x}}_1 - (\underline{A}_0 + \lambda \underline{A}_1) \underline{x}_1 \right] +$$
$$\varepsilon^3 \left[\underline{\dot{x}}_3 - (\underline{A}_0 + \lambda \underline{A}_1) \underline{x}_3 + \underline{f} (\underline{x}_1) \right] + \qquad (84)$$
$$\ldots$$

The linear solution is obtained by setting the terms of order ε equal to zero. This yields

$$\underline{\dot{x}}_1 = (\underline{A}_0 + \lambda \underline{A}_1) \underline{x}_1 \qquad (85)$$

Let λ_F be the value of λ for which one root crosses the imaginary axis and let

$$s = i \omega_F \qquad (86)$$

be the value of that root. Then the steady-state solution (i.e., the solution obtained by disregarding those terms of the solution related to roots which lie in the left-hand side of the complex s-plane) is

$$\underline{x}_1(t) = a \underline{\tilde{x}}_F e^{i\omega_F t} + a^* \underline{\tilde{x}}_F^* e^{-i\omega_F t} \qquad (87)$$

where $\underline{\tilde{x}}_F$ is the eigenvector of

$$\underline{A}_F = \underline{A}_0 + \lambda_F \underline{A}_1 \qquad (88)$$

i.e., the nontrivial solution of

$$\left[i\omega_F \underline{I} + \underline{A}_F \right] \underline{\tilde{x}}_F = 0 \qquad (89)$$

whereas a reduces to a constant as ε goes to zero. In order to study the nonlinear system set

$$\lambda = \lambda_F + \lambda_2 \varepsilon^2 + \cdots \qquad (90)$$

where $\lambda_2 = \pm 1$ and (in order to allow for the dependence of the frequency on λ) assume

$$a = a_0 e^{i \varepsilon^2 \omega_2 t} \qquad (91)$$

so that

$$\underline{x}_1(t) = a_0 \underline{\tilde{x}}_F e^{i\omega t} + a_0 \underline{\tilde{x}}_F^* e^{-i\omega t} \qquad (92)$$

where

$$\omega = \omega_F + \varepsilon^2 \omega_2 + \cdots \tag{93}$$

Substituting Equations 90 to 93 into Equation 84 yields

$$\varepsilon\left\{\left[i\omega_F \underline{I} - \underline{A}_0 - \lambda_F \underline{A}_1\right] \tilde{\underline{x}}_F \, a \, e^{i\omega_F t} + \left[-i\omega_F \underline{I} - \underline{A}_0 - \lambda_F \underline{A}_1\right] \tilde{\underline{x}}_F^* \, a^* e^{-i\omega_F t}\right\} +$$
$$\varepsilon^3\left\{ \dot{\underline{x}}_3 - (\underline{A}_0 + \lambda_F \underline{A}_1)\underline{x}_3 + (i\omega_2 \underline{I} - \lambda_2 \underline{A}_1)\tilde{\underline{x}}_F \, e^{i\omega_F t} + (-i\omega_2 \underline{I} - \lambda_2 \underline{A}_1)\tilde{\underline{x}}_F^* \, e^{-i\omega_F t} - \underline{f}(\underline{x}_1)\right\} + \cdots = 0 \tag{94}$$

or (using Equation 89 and setting the ε^3-terms equal to zero)

$$\dot{\underline{x}}_3 - (\underline{A}_0 + \lambda_F \underline{A}_1)\underline{x}_3 = (-i\omega_2 \underline{I} + \lambda_2 \underline{A}_1)\tilde{\underline{x}}_F \, a \, e^{i\omega_F t} + (i\omega_2 \underline{I} + \lambda_2 \underline{A}_1)\tilde{\underline{x}}_F^* \, a^* e^{-i\omega_F t} + \tilde{\underline{f}}_3 \, a^3 e^{i3\omega_F t} + \tilde{\underline{f}}_1 \, a^2 a^* e^{i\omega_F t} + \tilde{\underline{f}}_3^* \, a^{*3} e^{-i3\omega_F t} + \tilde{\underline{f}}_1^* \, a^{*2} a \, e^{-i\omega_F t} \tag{95}$$

where

$$\tilde{\underline{f}}_3 = \left\{\sum_i \sum_j \sum_k c_{hijk} \, x_{F_i} \, x_{F_j} \, x_{F_k}\right\} \tag{96}$$

$$\tilde{\underline{f}}_1 = \left\{\sum_i \sum_j \sum_k c_{hijk}\left(x_{F_i} \, x_{F_j} \, x_{F_k}^* + x_{F_i} \, x_{F_j}^* \, x_{F_k} + x_{F_i}^* \, x_{F_j} \, x_{F_k}\right)\right\} \tag{97}$$

A periodic solution exists if (and only if) the $e^{i\omega_F t}$-terms to not have a component in the $\tilde{\underline{x}}_F$-direction, i.e., if

$$\tilde{\underline{x}}_L^T \left[(-i\omega_2 \underline{I} + \lambda_2 \underline{A}_1) \, a \, \tilde{\underline{x}}_F + a^2 a^* \, \tilde{\underline{f}}_1\right] = 0 \tag{98}$$

where $\tilde{\underline{x}}_L^T$ is the left-eigenvector of $\underline{A}_F$, i.e., the nontrivial solution of

$$\tilde{\underline{x}}_L^T \left[i\omega_F \underline{I} - \underline{A}_F\right] = 0 \tag{99}$$

which is assumed to be normalized so that

$$\tilde{\underline{x}}_L^T \, \tilde{\underline{x}}_F = 1 \tag{100}$$

Equation 98 may be rewritten as

$$i\omega_2 a_0 + \beta a_0 + \gamma a_0^2 a_0^* \qquad (101)$$

where (using Equation 100)

$$\beta = -\lambda_2 \, \tilde{\underline{x}}_L^T \underline{A}_1 \tilde{\underline{x}}_F \qquad (102)$$

and

$$\gamma = -\tilde{\underline{x}}_L^T \tilde{\underline{f}}_1 \qquad (103)$$

Equation 101 is equal to Equation 33: the discussion of this equation is the same as that of Section V, and is summarized here.

Consider first the case in which the nonlinear-term stability parameter, γ_R, is positive: then the nonlinear terms are stabilizing and for $\lambda<\lambda_F$ the system is stable. On the other hand, for $\lambda>\lambda_F$ (where the linear analysis predicts instability), there exists a steady state periodic solution with finite amplitude (limit-cycle solution), given by

$$\underline{x}(t) = |\lambda-\lambda_F|^{1/2} \left|\frac{\underline{\beta_R}}{\gamma_R}\right|^{1/2} 2\,\mathrm{Real}\left(e^{i(\omega t+\varphi)} \tilde{\underline{q}}_F\right) + \cdots \qquad (104)$$

with

$$\omega = \omega_F + |\lambda-\lambda_F|\left(\gamma_I \beta_R/\gamma_R - \beta_I\right) + \cdots \qquad (105)$$

where β_R and γ_R (β_I and γ_I) are the real (imaginary) parts of β and γ. This behavior is referred to as benign flutter.

In the second case, the nonlinear-term stability parameter, γ_R, is negative: then the nonlinear terms are destabilizing and for $\lambda>\lambda_F$ the system is unstable (it goes to infinity in finite time: destructive flutter). On the other hand, for $\lambda>\lambda_F$ (the linear analysis predicts stability!), there exists a periodic solution with finite amplitude (also given by Equation 104); however, this solution is unstable (unstable limit-cycle solution) and it goes to zero (like the linear analysis predicts) only if the initial conditions are

such that the initial complex-amplitude of the $\tilde{q}_{I_F}$-component of $\underline{x}$ is less than its limit-cycle amplitude, $|\lambda-\lambda_F|^{1/2}|\beta_R/\gamma_R|^{1/2}$. Otherwise, the solution goes to infinity (in finite time): in other words, for $\gamma_R<0$, the system may experience destructive flutter even below the flutter speed, provided the initial complex-amplitude of the $\tilde{q}_{I_F}$-component of $\underline{x}$ is higher than the unstable limit-cycle amplitude (this could be caused, for instance, by gust response).

VIII. FLUTTER TAMING FOR AN AEROSERVOELASTIC SYSTEM

It is apparent from the remarks at the end of the preceding section that flutter is benign if the stability parameter of the nonlinear terms, γ_R, is positive. Therefore, flutter taming may be achieved by augmenting the system in such a way that the nonlinear-term-stability-parameter of the augmented system, γ_R^+, is positive. In order to accomplish this, consider the augmented system

$$\dot{\underline{x}} = (\underline{A}_0 + \lambda \underline{A}_1)\,\underline{x} + \underline{f}(\underline{x}) + \underline{B}\,\underline{u} + \underline{f}^{v}(\underline{x},\underline{u}) \tag{106}$$

$$\dot{\underline{u}} = \underline{D}\,\underline{u} + \alpha\,\underline{v}\,(\underline{c}^T\underline{x})^3 \tag{107}$$

[From a physical point of view this modification can be accomplished by having an additional control surface (the implementation is similar if one wants to use a control surface already used for flutter suppression): in this case

$$\underline{u} = \left\{ \begin{matrix} \delta \\ \cdots \\ \dot{\delta} \end{matrix} \right\} \tag{108}$$

where δ is the deflection of the control surface, $\underline{B}\,\underline{u}+\underline{f}^{v}$ is the vector of the aerodynamic forces caused by δ and $\dot{\delta}$, whereas $\underline{D}$ represents the dynamics of the control surface actuator (which is assumed not to be affected by the pressure acting on it) and the feedback term $\alpha\,\underline{v}(\underline{c}^T\underline{x})^3$ is the "driving force" for the actuator of the control surface, which is proportional to the third power of a linear combination of the state variables x_i. In particular, the system is equivalent to that considered in Section 5 if

$$\underline{D} = \begin{bmatrix} 0 & 1 \\ -a_0 & -a_1 \end{bmatrix} \tag{109}$$

$$\underline{v} = \left\{ \begin{matrix} 1 \\ 0 \end{matrix} \right\} \tag{110}$$

$$\underline{c} = \left\{ \begin{array}{c} c_0 \\ \hline c_1 \end{array} \right\} \tag{111}$$

and

$$\underline{\check{f}} = \left\{ \begin{array}{c} 0 \\ \hline \hat{f} \end{array} \right\} \tag{112}$$

It is important to emphasize that the derivation presented in the following is based on Equations 106 and 107 and therefore independent of the particular expressions for $\underline{D}, \underline{v}, \underline{\epsilon}$ and $\underline{\check{f}}$, i.e., independent of the physical implementation.]

Equations 106 and 107 may be recast as

$$\dot{\underline{x}}^+ = (\underline{A}_0^+ + \lambda \underline{A}_1^+)\, \underline{x}^+ + \underline{f}^+(\underline{x}^+) \tag{113}$$

where

$$\underline{x}^+ = \left\{ \begin{array}{c} \underline{x} \\ \hline \underline{u} \end{array} \right\} \tag{114}$$

$$\underline{A}_0^+ = \left[\begin{array}{c|c} \underline{A}_0 & \underline{0} \\ \hline \underline{0} & \underline{D} \end{array} \right] \tag{115}$$

$$\underline{A}_1^+ = \left[\begin{array}{c|c} \underline{A}_1 & \underline{B} \\ \hline \underline{0} & \underline{0} \end{array} \right] \tag{116}$$

$$\underline{f} = \left\{ \begin{array}{c} \underline{f} + \underline{\check{f}} \\ \hline \alpha \underline{v} \, (\underline{c}^T \underline{x})^3 \end{array} \right\} \tag{117}$$

Equation 113 is formally equal to Equation 82 and therefore the results of Section 6 are applicable to Equation 113 as well.

The solution of the linear system is obtained by setting

$$\det \left[s \underline{I} - \underline{A}_0^+ - \lambda \underline{A}_1^+ \right] =$$

$$\det \left[\begin{array}{c|c} s \underline{I} - \underline{A}_0 - \lambda \underline{A}_1 & -\lambda \underline{B} \\ \hline \underline{0} & s \underline{I} - \underline{D} \end{array} \right] =$$

$$\det\left[s\underline{I} - \underline{A}_o - \lambda \underline{A}_1\right] \det\left[s\underline{I} - \underline{D}\right] = 0 \qquad (118)$$

which shows that the characteristic roots of the augmented system are equal to those of the original system plus those of the actuator system.

In particular, for $\lambda = \lambda_F$, a root, $s = i\omega$, is on the imaginary axis and the corresponding eigenvector $\tilde{\underline{x}}_F$, i.e., a nontrivial solution of

$$\left[i\omega_F \underline{I} - \underline{A}_o^+ - \lambda_F \underline{A}_1^+\right] \tilde{\underline{x}}_F^+$$

$$= \left[\begin{array}{c|c} i\omega_F \underline{I} - \underline{A}_o - \lambda_F \underline{A}_1 & -\lambda_F \underline{B} \\ \hline \underline{0} & i\omega_F \underline{I} - \underline{D} \end{array}\right] \tilde{\underline{x}}_F^+ = 0 \qquad (119)$$

is given by

$$\tilde{\underline{x}}_F^+ = \left\{\begin{array}{c} \tilde{\underline{x}}_F \\ \hline \underline{0} \end{array}\right\} \qquad (120)$$

In addition, the left eigenvector $\tilde{\underline{x}}_L^+$, i.e., the solution of

$$\left[i\omega_F \underline{I} - \underline{A}_o^+ - \lambda_F \underline{A}_1^+\right] \tilde{\underline{x}}_L^+$$

$$= \left[\begin{array}{c|c} i\omega_F \underline{I} - \underline{A}_o^T - \lambda_F \underline{A}_1^T & \underline{0} \\ \hline -\lambda_F \underline{B}^T & i\omega_F \underline{I} - \underline{D}^T \end{array}\right] \tilde{\underline{x}}_L^+ = 0 \qquad (121)$$

is given by

$$\tilde{\underline{x}}_L^+ = \left\{\begin{array}{c} \tilde{\underline{x}}_L \\ \hline \tilde{\underline{u}}_L \end{array}\right\} \qquad (122)$$

where

$$\tilde{\underline{u}}_L = +\lambda_F \left[i\omega_F \underline{I} - \underline{D}^T\right]^{-1} \underline{B}^T \tilde{\underline{x}}_L \qquad (123)$$

The benign flutter condition is given by

$$\gamma_R^+ = \text{Real}\left(\tilde{\underline{x}}_L^{+T} \tilde{\underline{f}}_1^+\right) \qquad (124)$$

where

$$\underline{\tilde{f}}_1^{+} = \left\{ \frac{\underline{\tilde{f}}_1 + \underline{\tilde{\tilde{f}}}_1}{\alpha \underline{v}\, 3\mu^2\mu^*} \right\} \tag{125}$$

with $\underline{\tilde{f}}_1$ given by Equation 97, and $\underline{\tilde{\tilde{f}}}_1$ given by a similar expression, whereas

$$\mu = \underline{c}^T \underline{\tilde{x}}_F \tag{126}$$

Using Equations 122 and 125, Equation 124 may be rewritten as

$$\gamma_R^{+} = \gamma_{R_0} + \alpha \gamma_{R_1} > 0 \tag{127}$$

where

$$\gamma_{R_0} = Real\left[\underline{\tilde{x}}_L^T \left(\underline{\tilde{f}}_1 + \underline{\tilde{\tilde{f}}}_1 \right) \right] = Real\left(\gamma + \gamma^{v} \right) \tag{128}$$

(with $\gamma = \underline{\tilde{x}}_L^T \underline{\tilde{f}}_1$ (see Equation 103) due to the nonlinearities of the original system and $\gamma^{v} = \underline{\tilde{x}}_L^T \underline{\tilde{\tilde{f}}}_1$ due to the additional nonlinearities of the control surface aerodynamics) and

$$\gamma_{R_1} = 3|\mu|^2 Real\left(\mu\, \underline{\tilde{x}}_L^T \underline{v} \right) \tag{129}$$

By choosing $\underline{v}$ and $\underline{e}$ such that

$$|\mu| = |c^T \tilde{x}_F| \neq 0 \tag{130}$$

and

$$Real\left(\mu\, \underline{\tilde{x}}_L^T \underline{v} \right) = Real\left[\left(c^T \tilde{x}_F \right) \left(\underline{\tilde{x}}_L^T \underline{v} \right) \right] \neq 0 \tag{131}$$

one obtains

$$\gamma_{1R} \neq 0 \tag{132}$$

and hence it is always possible to find a value for α such that the benign flutter condition, Equation 127, is always satisfied.

Finally, as $\alpha \, sign(\gamma_{R_1})$ goes to infinity, γ_R^{+} goes to infinity and hence the limit-cycle amplitude goes to zero. In other words, the amplitude of the limit-cycle can be made as small as desired by choosing an appropriate value for the gain α .

IV. RECOMMENDATIONS

A theory for nonlinear flutter of aircraft has been presented. Four main assumptions have been used in the formulation.

The first one is that the linear portion of the aerodynamics forces can be expressed (in the time domain) as a linear combination of the generalized Lagrangian coordinates and of their derivative (in the frequency domain this corresponds to a complex aerodynamic matrix in which the real part is frequency independent and the imaginary part is proportional to the frequency).

The second one is that all the nonlinear terms be of algebraic nature (the assumption that second order nonlinear terms are equal to zero is not essential to the derivation of the solution; see References 11 and 15).

The third one is that terms of order higher than $|\lambda - \lambda_F|^{3/2}$ (where λ is the dynamic pressure and λ_F is its value at the flutter boundary) are negligible.

The last one is that the terms in the solution corresponding to characteristic roots with negative real parts do not yield any instability (it should be emphasized that, if the component of the nonlinear-term vector in the direction of any characteristic vector is negative, destructive-flutter response may occur for the corresponding component of the solution).

The solution indicates that within the above restrictive assumptions two phenomena (benign and destructive flutter) are possible, depending upon

the value of a number, γ_R , called the stability parameter of the nonlinear terms. If $\gamma_R > 0$ ($\gamma_R < 0$), then the nonlinear terms are stabilizing and benign (destructive) flutter occurs.

This result suggests that nonlinear active control may be used to obtain an augmented system which is such that its nonlinear terms are stabilizing, whereby, ensuring that only benign flutter occurs. It is shown that a remarkably simple nonlinear control law exists such that flutter taming (i.e., transforming a destructive-flutter system into a benign-flutter system) is always possible: this is accomplished by setting the magnitude of the gain (with appropriate sign) above a given value. In addition, the magnitude of the amplitude of the steady-stated (limit-cycle) amplitude can be controlled (as the magnitude of the gain goes to infinity the amplitude of the limit-cycle goes to zero).

It is recommended that this work be continued to include, in order, the following items:

1. An extension of the formulation to remove the four restrictive assumptions mentioned above.

2. A numerical validation of the theory by direct integration in the time domain of the equation governing the phenomenon .

3. A wind-tunnel verification of the results of the numerical validation (item 2). This can be accomplished by adding the nonlinear control law for flutter taming to an existing wind-tunnel model developed for the study of flutter suppression.

4. A flight test program for the actual implementation of the concept (provided, of course, that the results obtained under items 1, 2, and 3 confirm the results presented here).

APPENDIX A

LAGRANGIAN FORMATION OF MOTION

Let the displacement $\vec{\delta}$ of an aircraft be represented in terms of its natural modes of vibration, $\vec{M}_n$, (which form a closed and complete system of orthogonal functions),

$$\vec{\delta}(\xi^\alpha, t) = \sum_{n=1}^{N} q_n(t)\, \vec{M}_n(\xi^\alpha) \tag{A.1}$$

where t indicates time, ξ^α indicates a set of material (or convected) coordinates. The amplitudes, q_n, of the natural modes of vibrations are the generalized Lagrangian coordinates of the system and satisfy the Lagrangian equations of motion

$$\frac{d}{dt}\frac{\partial T}{\partial \dot{q}_n} - \frac{\partial T}{\partial q_n} - \frac{\partial U}{\partial q_n} = Q_n \tag{A.2}$$

where T is the kinetic energy and U is the elastic energy.

In addition, the generalized forces $Q_n = e_n + Q_n^D$ include the generalized aerodynamic forces

$$e_n = -\lambda \oint\!\!\!\oint_\sigma c_p\, \vec{n} \cdot \vec{M}_n\, d\sigma \tag{A.3}$$

(where $\lambda = \frac{1}{2}\rho_\infty V_\infty^2$ is the dynamic pressure and c_p is the aerodynamic pressure coefficients) as well as the effect of structural damping, Q_n^D.

Note that

$$T = \tfrac{1}{2}\iiint \rho v^2\, dV = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} M_{ij}\, \dot{q}_i\, \dot{q}_j \tag{A.4}$$

with (because of the orthogonality relations of the natural modes of vibration)

(A.5)

$$M_{ij} = \iiint_V \rho \vec{M}_i \cdot \vec{M}_j \, dV$$

$$= \iiint_V \rho \vec{M}_i \cdot \vec{M}_j \, |J| \, d\xi^1 d\xi^2 d\xi^3 = \begin{cases} m_i & (i = j) \\ 0 & (i \neq j) \end{cases}$$

where m_i are the generalized masses (note that $\rho|J|$ = constant because of conservation of mass: therefore, the generalized masses are independent of the deformation). For simplicity, in the following, the modes are assumed to be normalized so that

$$m_i = 1 \qquad (i = 1, 2 \dots) \tag{A.6}$$

In addition, by definition of natural modes of vibration,

$$\frac{\partial U}{\partial q_i} = -\omega_i^2 q_i + f_i^{(s)}(q_1, q_2, \dots) \tag{A.7}$$

where ω_i^2 are the natural frequencies and f_i are higher-order nonlinear terms (functions of $q_1, q_2 \dots$) which are assumed to be of algebraic nature, i.e.,

$$f_h^{(s)} = \sum_{i=1}^{N} \sum_{j=1}^{N} b_{hij}^{(s)} q_i q_j + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} c_{hijk}^{(s)} q_i q_j q_k + \dots \tag{A.8}$$

For simplicity, it is assumed that the aerodynamic forces can be approximated as

$$e_h = \lambda \left(\sum_i e_{hi}^{(0)} q_i + \sum e_{hi}^{(1)} \dot{q}_i + f_h^{(A)}(q_1, q_2, \dots, \dot{q}_1, \dot{q}_2, \dots) \right) \tag{A.9}$$

where $f_h^{(A)}$ are higher-order nonlinear aerodynamic terms also of algebraic nature, i.e.,

$$f_h^{(A)} = \sum_{i=1}^{2N} \sum_{j=1}^{2N} b_{hij}^{(A)} x_i x_j + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} c_{hijk}^{(A)} x_i x_j x_k + \cdots \quad (A.10)$$

where

$$\{x_i\} = \begin{Bmatrix} q_i \\ \dot{q}_i \end{Bmatrix} \quad (A.11)$$

Finally, the structural damping forces are approximated as

$$Q_h^{(D)} = \sum_{i=1}^{2N} r_{hi} \dot{q}_i + \sum_{i=1}^{2N} \sum_{j=1}^{2N} b_{hij}^{(D)} x_i x_j + \sum_{i=1}^{2N} \sum_{j=1}^{2N} \sum_{k=1}^{2N} c_{hijk}^{(D)} x_i x_j x_k + \cdots \quad (A.12)$$

Combining the above equations one obtains

$$\ddot{q}_h + \sum_{i=1}^{N} r_{hi} \dot{q}_i + \omega_h^2 q_h + \lambda \sum_{i=1}^{N} \left(e_{hi}^{(0)} q_i + e_{hi}^{(1)} \dot{q}_i \right) = f_h(q_i, \dot{q}_i) \quad (A.13)$$

where

$$f_h = \sum_{i=1}^{2N} \sum_{i=1}^{2N} b_{hij} x_i x_j + \sum_{i=1}^{2N} \sum_{i=1}^{2N} \sum_{i=1}^{2N} c_{hijk} x_i x_j x_k + \cdots \quad (A.14)$$

with

$$b_{hij} = b_{hij}^{(S)} + b_{hij}^{(A)} + b_{hij}^{(D)}$$
$$c_{hijk} = c_{hijk}^{(S)} + c_{hijk}^{(A)} + c_{hijk}^{(D)} \quad (A.15)$$

(where $b_{hij}^{(S)}$ and $c_{hijk}^{(S)}$ are understood to be equal to zero for i, or j, or K greater than N)

Equation A.10 may be recast in a form which is convenient for the analysis of Section 7. Introducing

$$\dot{q}_h = v_h \tag{A.16}$$

Equation A.10 may be rewritten as

$$\dot{v}_h + \sum_{i=1}^{N} r_{hi} v_i + \omega_h^2 q_h + \lambda \Sigma \left(e_{hi}^{(0)} q_i + e_{hi}^{(1)} \dot{q}_i \right) = f_h \tag{A.17}$$

The above two equations may be combined to yield

$$\begin{Bmatrix} \dot{q}_h \\ \dot{v}_h \end{Bmatrix} = \begin{bmatrix} 0 & \delta_{hi} \\ -K_{hi} & -r_{hi} \end{bmatrix} \begin{Bmatrix} q_i \\ v_i \end{Bmatrix} + \lambda \begin{bmatrix} 0 & 0 \\ -e_{hi}^{(0)} & -e_{hi}^{(1)} \end{bmatrix} \begin{Bmatrix} q_i \\ v_i \end{Bmatrix} + \begin{Bmatrix} 0 \\ f_h \end{Bmatrix} \tag{A.18}$$

where

$$K_{hi} = \omega_i^2 \delta_{hi} \tag{A.19}$$

or,

$$\underline{\dot{x}} = (\underline{A}_0 + \underline{\lambda}\, \underline{A}_1)\, \underline{x} + \underline{f} \tag{A.20}$$

where

$$\underline{x} = \begin{Bmatrix} q_h \\ v_h \end{Bmatrix}$$

$$\underline{A}_0 = \begin{bmatrix} 0 & \delta_{hi} \\ -K_{hi} & -r_{hi} \end{bmatrix}$$

$$\underline{A}_1 = \begin{bmatrix} 0 & 0 \\ - & - \\ -e_{hi}^{(0)} & -e_{hi}^{(1)} \end{bmatrix}$$

$$\underline{f} = \begin{Bmatrix} 0 \\ - \\ f_h \end{Bmatrix}$$

(A.21)

For analysis presented in Section 5 it is necessary to recast the equations in a different form: this is possible if the nonlinear effects in the structural damping terms are negligible and the aerodynamic nonlinear terms are functions of q_i but not functions of $\dot{q}_i$: in this case

$$b_{hij} = c_{hijk} = 0 \qquad \text{(A.22)}$$

if i, or j, or k is greater than N, and hence, f_h is not function of $\dot{q}_i$ so that

$$f_h = \sum_{i=1}^{N}\sum_{j=1}^{N} b_{hij}\, q_i\, q_j + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} c_{hijk}\, q_i\, q_j\, q_k + \cdots \qquad \text{(A.23)}$$

Then Equation A.10 may be rewritten as

$$\ddot{\underline{q}} + \underline{R}\,\dot{\underline{q}} + \underline{k}\,\underline{q} + \lambda\left(\underline{E}_0\,\underline{q} + \underline{E}_1\,\dot{\underline{q}}\right) = \underline{f}(\underline{q}) \qquad \text{(A.24)}$$

where

$$\underline{R} = [r_{hi}]$$
$$\underline{K} = [\omega_h^2]$$
$$\underline{E}_0 = [e_{hi}^{(0)}]$$
$$\underline{E}_1 = [e_{hi}^{(1)}] \qquad \text{(A.25)}$$

APPENDIX B

TWO TIME SCALE ANALYSIS

Consider a system of the type

$$\dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{f}(\underline{x}) \tag{B.1}$$

where $f(x)$ is the vector of all the nonlinear terms which are assumed to be of algebraic nature. Assume that

$$\underline{A} = \underline{A}_0 + \underline{\lambda}\,\underline{A}_1 \tag{B.2}$$

and that all the eigenvalues of $\underline{A}$ have negative real part for $\lambda < \lambda_F$ and that one root crosses the imaginary axis (not at the origin) at $\lambda = \lambda_F$ (stability boundary. Let $i\omega_F$ be that root. Note that all the roots in the left hand side of the plane yield terms that go to zero as t goes to infinity: therefore for $\lambda = \lambda_F$ the steady-state solution of the linear problem ($\underline{f} = 0$) is

$$\underline{x}(t) = c\,\underline{\tilde{x}}_F\, e^{i\omega_F t} + c^*\,\underline{\tilde{x}}_F^*\, e^{-i\omega_F t} \tag{B.3}$$

where c is a constant and $\underline{\tilde{x}}_F$ is the eigenvector of

$$\underline{A}_F = \underline{A}_0 + \lambda_F \underline{A}_1 \tag{B.4}$$

corresponding to the eigenvalue $i\omega_F$, i.e., the nontrivial solution of

$$(i\omega_F \underline{I} + \underline{A}_F)\,\underline{\tilde{x}}_F = 0 \tag{B.5}$$

Assume for simplicity that the second-order nonlinear terms are equal to zero (see Reference 11 and 14 for the case of nonzero second-order nonlinear terms). An approximate solution of Equation B.1 can be obtained as follows. Let

$$\varepsilon = |\lambda - \lambda_F| \tag{B.6}$$

so that

$$\lambda = \lambda_F + \lambda_2\,\varepsilon^2 \tag{B.7}$$

with

$$\lambda_2 = \pm 1 \tag{B.8}$$

and assume that the solution for $\underline{x}$ is function of time as follows:

$$\underline{x} = \underline{x}(t, \tau) \tag{B.9}$$

where

$$\tau = \varepsilon^2 t \tag{B.10}$$

This means that time dependence appears in two different forms: the first one is independent of ε while the second one, the so-called slow-scale, becomes slower as ε goes to zero. On the other hand, introducing the transformation given by Equation B.10, the dependence of $\underline{x}$ upon τ is independent of ε, whereas, the dependence upon the fast scale τ/ε^2 becomes faster and faster as ε goes to zero.

Equation B.9 implies that

$$\dot{\underline{x}} = \frac{d}{dt}\underline{x} = \left(\frac{\partial}{\partial t} + \varepsilon^2 \frac{\partial}{\partial \tau}\right)\underline{x} \tag{B.11}$$

Consider a solution of the type

$$\underline{x} = \varepsilon \underline{x}_1 + \varepsilon^3 \underline{x}_3 + \cdots \tag{B.12}$$

where $\underline{x}_i$ are functions of t and τ.

Combining Equations B.1, B.2, B.4, B.7, B.11 and B.12 yields

$$\begin{aligned} &\varepsilon\left[\frac{\partial \underline{x}_1}{\partial t} - \underline{A}_F \underline{x}_1\right] + \\ &\varepsilon^3\left[\frac{\partial \underline{x}_3}{\partial t} - \underline{A}_F \underline{x}_3 + \frac{\partial \underline{x}_1}{\partial \tau} - \lambda_2 \underline{A}_1 \underline{x}_1 - \underline{f}(x_1)\right] + \\ &\cdots = 0 \end{aligned} \tag{B.13}$$

where all the terms in brackets are not explicitly function of ε (it may be noted that this is the mathematical reason for introducing the slow time scale τ). Separating terms of the same order of magnitude yields

$$\frac{\partial \underline{x}_1}{\partial t} - \underline{A}_F \underline{x}_1 = 0 \tag{B.14}$$

$$\frac{\partial \underline{x}_3}{\partial t} - \underline{A}_F \underline{x}_3 = -\frac{\partial \underline{x}_1}{\partial \tau} + \lambda_2 \underline{A}_1 \underline{x}_1 + \underline{f}(\underline{x}_1) \qquad \text{(B.15)}$$

Disregarding the terms in the solution corresponding to characteristic roots with negative real parts, the solution of Equation B.14 is

$$\underline{x}_1 = a(\tau)\, \tilde{\underline{x}}_F\, e^{i\omega_F t} + a^*(\tau)\, \tilde{\underline{x}}_F^*\, e^{-i\omega_F t} \qquad \text{(B.16)}$$

where $i\omega_F$ is the imaginary eigenvalue of $\underline{A}_F$ (see Equation B.4) and $\tilde{\underline{x}}_F$ is the corresponding eigenvector (see Equation B.5): the main difference between the solution of the linear problem, Equation B.3, and the first order solution of the nonlinear problem, Equation B.16, is that, in general, a is not a constant but a function of τ (note that in Equation B.14 the time derivative is a partial derivative, whereas in the linear problem (Equation B.1 with $\underline{f} = 0$) the time derivative is an ordinary derivative).

Substituting Equation B.16 into Equation B.15, and noting that

$$\begin{aligned}\underline{f}(\underline{x}_1) = &\Big\{\sum_i \sum_j \sum_k \Big[c_{ijk} \left(a x_{F_i} e^{i\omega_F t} + a^* x_{F_i}^* e^{-i\omega_F t}\right) \times \\ &\left(a x_{F_j} e^{i\omega_F t} + a^* x_{F_j}^* e^{-i\omega_F t}\right) \times \\ &\left(a x_{F_k} e^{i\omega_F t} + a^* x_{F_k}^* e^{-i\omega_F t}\right)\Big]\Big\} \\ = &\tilde{\underline{f}}_3\, a^3 e^{i3\omega_F t} + \tilde{\underline{f}}_3^*\, a^{*3} e^{-i3\omega_F t} + \\ &\tilde{\underline{f}}_1\, a^2 a^* e^{i\omega_F t} + \tilde{\underline{f}}_1^*\, a^{*2} a\, e^{-i\omega_F t}\end{aligned} \qquad \text{(B.17)}$$

one obtains

$$\begin{aligned}\frac{\partial \underline{x}_3}{\partial t} - \underline{A}_F \underline{x}_3 = &\; a^3 \tilde{\underline{f}}_3\, e^{i3\omega_F t} + \\ &\left[-\tilde{\underline{x}}_F \frac{\partial a}{\partial \tau} + \lambda_2 \underline{A}_1 \tilde{\underline{x}}_F a + \tilde{\underline{f}}_1 a^2 a^*\right] e^{i\omega_F t} \\ &+ \text{conjugate terms}\end{aligned} \qquad \text{(B.18)}$$

The solution of this equation contain "secular terms" (i.e., terms of the type $t e^{i\omega_F t}$) unless

$$\tilde{\underline{X}}_L^T \left[-\tilde{\underline{X}}_F \frac{\partial a}{\partial t} + \lambda_2 \underline{A}_1 \tilde{\underline{X}}_F a + \tilde{\underline{f}}_1 a^2 a^* \right] = 0 \quad \text{(B.19)}$$

where $\tilde{\underline{X}}_L$ is the left eigenvector of $\underline{A}_F$ corresponding to the eigenvalue $i\omega_F$, i.e., the nontrivial solution of

$$\left[i\omega_F \underline{I} - \underline{A}_F \right]^T \tilde{\underline{X}}_L = 0 \quad \text{(B.20)}$$

If $\tilde{\underline{X}}_L$ is normalized so that

$$\tilde{\underline{X}}_L^T \tilde{\underline{X}}_F = 1 \quad \text{(B.21)}$$

Equation B.19 may be rewritten as

$$\frac{\partial a}{\partial \tau} + \beta a + \gamma a^2 a^* = 0 \quad \text{(B.22)}$$

where

$$\beta = -\lambda_2 \tilde{\underline{X}}_L^T \underline{A}_1 \tilde{\underline{X}}_F \quad \text{(B.23)}$$

$$\gamma = -\tilde{\underline{X}}_L^T \tilde{\underline{f}}_1 \quad \text{(B.24)}$$

By setting

$$a = |a| e^{i\varphi}$$

and separating real and imaginary parts yields

$$\frac{\partial |a|}{\partial \tau} + \beta_R |a| + \gamma_R |a|^3 = 0 \quad \text{(B.25)}$$

$$\frac{\partial \varphi}{\partial t} + \beta_I + \gamma_I |a|^2 = 0 \quad \text{(B.26)}$$

Equation B.25 may be solved by setting $|a| = \frac{1}{\sqrt{z}}$ to obtain

$$\frac{\partial z}{\partial \tau} - 2\beta_R z - 2\gamma_R = 0 \quad \text{(B.27)}$$

or

$$z = -\frac{\gamma_R}{\beta_R} \left(1 + k e^{2\beta_R \tau} \right) \quad \text{(B.28)}$$

(where k depends upon the initial condition) and hence

$$|a| = \left| \frac{-\beta_R / \gamma_R}{1 + k e^{2\beta_R \tau}} \right| \quad \text{(B.29)}$$

Substituting into Equation B.26 and integrating yields

$$\varphi = -\int_0^\tau (\beta_I + \gamma_I |a|^2)\, d\tau + \varphi_0$$

$$= (\gamma_I \beta_R / \gamma_R - \beta_I)\,\tau + \frac{\gamma_I}{\gamma_R} \ln|a| + \varphi_0 \qquad \text{(B.30)}$$

Note that

$$\beta_R \lesseqgtr 0 \quad \text{for} \quad \lambda_2 = \pm 1 \quad (\lambda \gtrless \lambda_F) \qquad \text{(B.31)}$$

because, by definition of λ_F, the effect of the nonlinear terms is destabilizing (stabilizing) for $\lambda > \lambda_F$ ($\lambda < \lambda_F$).

REFERENCES

1. Bisplinghoff, R.L., Ashley, H., Hoffman, R.L., Aeroelasticity, Addison-Wesley Publishing Company, Inc., 1955.

2. Garrick, I.E., "Aeroelasticity-Frontiers and Beyond, 13th von Karman Lecture", Journal of Aircraft, Vol. 13, No. 9, Sep 1976, pp. 641-657.

3. Edwards, J.W., "Unsteady Aerodynamic Modelling and Active Aeroelastic Control", Stanford University Center for Systems Research, SUDAAR 504, Feb 77.

4. Librescu, L. "Aeroelastic Stability of Orthotropic Heterogeneous Thin Panels in the Vicinity of the Flutter Critical Boundary, Part I", J. Mechanique IV, 1965, pp. 51-76.

5. Kobayashi, S., "Flutter of Simply Supported Rectangular Panels in a Supersonic Flow - Two-Dimensional Panel Flutter, I Simply Supported Panels, II Clamped Panel", Trans. Japan Soc. Aeronaut. Sci., Vol. 5, pp. 79-118.

6. Dowell, E.H., "Nonlinear Oscillation of a Fluttering Plate", AIAA J., Vol. 4, 1966, pp. 1267-1275.

7. Dowell, E.H., "Nonlinear Oscillation of a Fluttering Plate II", AIAA J., Vol. 5, 1967, pp. 1856-1862.

8. Dowell, E.H., "Panel Flutter: A Review of the Aeroelastic Stability of Plates and Shells", AIAA J., Vol. 8, No. 3, March 1970, pp. 385-399.

9. Dowell, E.H., Aeroelasticity of Plates and Shells, Noordhoff International Publishing, Layden, 1975.

10. Dowell, E.H., et al., A Modern Course in Aeroelasticity, Sijthoff and Noordhoff, 1978.

11. Morino, L. "A Perturbation Method for Treating Nonlinear Panel Flutter Problems", AIAA J. Vol. 7, 1969, pp. 405-411.

12. Sandri, G., Phys. Rev. Letters, Vol. 11, p. 178 (1963); Ann. Phys., Vol 24, pp. 332-380.

13. Frieman, E.A., J. Math. Physics, Vol. 4, p. 410 (1963).

14. Nayfeh, J.A., "A Perturbation Method for Treating Nonlinear Oscillation Problems", Journal of Mathematics and Physics, Vol. XLIV, 1965, pp. 368-374.

15. Smith, L.L. and Morino, L., "Stability Analysis of Nonlinear Differential Autonomous Systems with Applications to Flutter", AIAA J., Vol. 14, 1976, pp. 333-341.

16. Morino, L. and Kuo, C.C., "Detailed Extension of Perturbation Method for Panel Flutter", ASRL TR-164-2, March 1971, Aeroelastic and Structures

Research Lab., Dept. of Aeronautics and Astronautics, Massachusetts Institude of Technology, Cambridge, Mass.

17. Davis, W.H., Jr., "Nonlinear Flutter of Panels and Wings by Lie Transformation Perturbation Method", Boston University, Dept. of Aerospace Engineering, TR-73-04, 1973.

18. Cole, J. and Kevorkian, J., "Uniformly Valid Asymptotic Approximation for Certain Nonlinear Differential Equations", Nonlinear Differential Equations in Nonlinear Mechanics, Academic Press, New York, Vol. 9, 1963, pp. 113-120.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

A THEORY OF PHOTOLUMINESCENCE FOR EXCITONS BOUND TO TWO TYPES OF NEUTRAL ACCEPTORS IN SILICON - A STUDY OF THE SYSTEMS Si: (In,Al) AND Si: (In,B).

Prepared by:	Dr. David S. Moroi
Academic Rank:	Professor
Department and University:	Department of Physics Kent State University
Research Location:	Air Force Wright Aeronautical Laboratories, Materials Laboratory, Electromagnetic Materials Division, Laser and Optical Materials Branch
USAF Research Colleagues:	Mr. David H. Brown, Dr. Patrick M. Hemenger, and Dr. Melvin C. Ohmer
Date:	August 12, 1982
Contract No.:	F49620-82-C-0035

A THEORY OF PHOTOLUMINESCENCE FOR EXCITONS BOUND TO TWO TYPES OF NEUTRAL ACCEPTORS IN SILICON- A STUDY OF THE SYSTEMS Si: (In, Aℓ) AND Si: (In, B)

by

David S. Moroi

ABSTRACT

In the absence of saturation, the rate equations for the densities of free excitons and excitons bound to two types of neutral acceptors in silicon are solved for steady state. The rate equations contain the terms for the transfer (tunneling) of an exciton bound to one type of neutral impurity to another and its reverse transfer. The steady state solutions of the rate equations yield an expression for the ratio of the bound exciton luminescence intensities as a function of the impurity concentrations. The rate of transfer for a bound exciton is calculated using the standard treatment in quantum mechanics with the help of a hydrogenic model wavefunction for an exciton bound to a neutral acceptor. This model exciton-neutral-acceptor complex consists of two relatively massive holes attracted to the negative acceptor center forming a core and a single relatively light electron repelled by the negative acceptor center. The bound exciton transfer rate is expressed in terms of some forty two exponential integrals, more than one half of which can be obtained using the recurrence relations among the basic two-center integrals. The transfer rate is also calculated using the WKB approximation in quantum mechanics which resulted the same distance-dependent major factor.

Acknowledgement

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to spend a very worthwhile and rewarding summer at the Materials Laboratory, Wright-Patterson AFB, Ohio. He would like to extend his appreciation to the laboratory, in particular the Laser and Optical Materials Branch, for their hospitality and continuing support during his tenured term.

Finally, he would like to thank Dr. Patrick M. Hemenger and Dr. Melvin C. Ohmer for suggesting this area of research and their collaboration and guidance, and, above all, he would like to express his deep gratitude to Mr. David H. Brown for numerous helpful discussions and for providing many useful data and references, and to Dr. Gust Bambakidis, a SCEEE fellow, and Dr. Frank Szmulowicz for several stimulating discussions.

I. INTRODUCTION:

Characterization of semiconductor materials in terms of their electromagnetic properties is important in the development of semiconductor devices. Various electric and optical techniques have been ultilized to study the electromagnetic properties of semiconductors.[1] Particularly the recent progress in technology of integrated circuits demands near-perfect crystals of silicon and, therefore, a great improvement in the methods of analysis for minute amount of impurities and defects in silicon is required, since they drastically change its electromagnetic properties.

Recently attempts have been made to use photoluminescence (PL) as a quantitative tool for measuring impurity concentrations. Mitchard and McGill[2] used PL to determine relative concentrations in the Si: (B,In) system. Brown et.al.[3] investigated the optical and electrical characterization of the system Si: (In, Aℓ) using PL, Hall effect transport, infrared absorption and photoconductivity measurements. In both experiments, they observed that the luminescence from the shallower acceptor (Aℓ,B) of a two acceptor system may be quenched when the concentration of the deeper acceptor (In) exceeds some initial values. [2-4] In other words, no aluminum nor boron bound exciton luminescence was observed despite high concentration of Aℓ and B. It is believed that the tunneling of bound exciton (BE) from Aℓ(B) to In bound states is responsible for the quenching of the Aℓ(B) luminescence.

In order to study this phenomena we attempted to derive an equation

$$I_1/I_2 = R_{12}\, N_1/N_2, \tag{1}$$

where I_1 and I_2 are the PL intensities for the Aℓ(B) BE and the In BE, N_1 and N_2 the concentrations of Aℓ(B) and In, respectively and the factor R_{12} includes the exciton transfer and its reverse transfer rates, the BE decay rates, the BE oscillator strengths etc. From here on, we will use the subscript "$_1$" and "$_2$" for quantities associated with an Aℓ(B) atom and with an In atom, respectively; e.g. N_1 = the concentration of Aℓ(B) atoms, I_2 = the intensity of PL from the In BE. The rate equations[5] for the densities of the free excitons (FE) and of the Aℓ(B) and In BE including the BE transfer from an Aℓ(B) atom to a nearest neighbor In atom and its reverse

transfer are solved in the steady state and in the absence of saturation. The exciton transfer rate is calculated using the standard method in quantum mechanics with the help of a model wavefunction for an exciton bound to a neutral acceptor which consists of two relatively massive holes attracted to the negative acceptor center forming a core and a single relatively light electron repelled by the negative acceptor center.[6] The tunneling rate is also calculated using the WKB method in a semi-classical approximation in quantum mechanics.[7] It is amusing to know that both calculations give the same distance-dependence in the major factor of the exciton transfer rate.

II. OBJECTIVES

The main goals and objectives of this research project were to understand the electromagnetic properties and structure of multiply doped silicon, and, in particular, to provide a theory interpreting the experimental results on PL from the Si: (In,Aℓ) system by Brown et al.[3] We did not attempt to solve the general problem because of its complexity. Instead, we treated the simplest case, leaving more complicated cases for later investigations. Our specific goals and objectives were:

(1) To derive Eq. (1) which gives the relationship between the PL intensity ratio and the impurity concentrations.

(2) To calculate the BE tunneling and its reverse tunneling rates, using reasonable wavefunctions for the Aℓ(B) and the In BE.

(3) To compute the integrals involving the exciton transfer rates.

III. THE RATE MODEL OF DOPED Si PHOTOLUMINESCENCE INCLUDING EXCITON TRANSFER

A. Impurity Distribution

To study the effect on luminescence of increasing Aℓ(B) and In concentrations, we consider here some direct interaction between an Aℓ(B)BE and its nearest-neighbor neutral In atom which results in a transfer of the exciton from the Aℓ(B) atom to the In atom (the overlapping of the wavefuncitons for the Aℓ(B)BE and In BE).

We calculate first the probability that the nearest-neighbor In atom is between r and r + dr from a given Aℓ(B) atom as shown in Fig. 1, $P_2(r)dr$:

$$P_2(r)dr = 4\pi r^2 dr N_2 \exp(-4\pi/3\ r^3 N_2) \tag{2}$$

where

$4\pi r^2 dr N_2$ = The probability that an In atom is between r and r + dr from a given Aℓ(B) atom.

$\exp(-4\pi/3\ r^3 N_2)$ = The probability that there is no In atom between the Aℓ(B) atom r (obtained from the Poisson distribution).

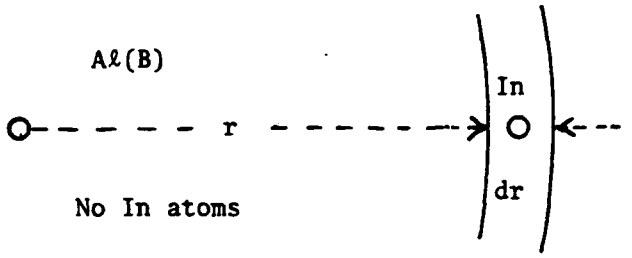


FIGURE 1 - CONFIGURATION OF IMPURITY ATOMS

Note that

$$\int_0^\infty P_2(r)dr = 1. \tag{3}$$

The average value of r for a given In concentration is

$$\langle r\rangle_2 = \int_0^\infty rP_2(r)dr = (3/4\pi N_2)^{1/3}\ \Gamma(4/3) \tag{4}$$

For $N_2=10^{16}cm^{-3}$ we find $\langle r\rangle_2$ = 250Å. Considering that Si BE Bohr radius is 43Å, it is reasonable to assume that the direct interaction effects become important.

Now the density of impurity Aℓ(B) atoms with a nearest-neighbor In between r and r + dr is given by

$$N_1(r) = N_1 P_2(r), \tag{5}$$

so that

$$\int_0^\infty N_1(r)dr = N_1 \int_0^\infty P_2(r)dr = N_1 \tag{6}$$

Simply switching the subscript "2" to subscript "1" in Eqs. (2) and (5), we obtain the probability and the density of In atoms that the nearest-neighbor Aℓ(B) atom is between r and r + dr from a given In atom.

B. The Rate Equations

In order to derive Eq. (1) on the basis of the model for the formation and decay kinetics of FE and BE including the BE transfer and its reverse transver, we start with the rate equations for the densities of the FE and BE.

$$\frac{dn_{FE}}{dt} = g - (\nu_{FE} + \gamma_1 + \gamma_2)n_{FE} + \rho_1 n_1 + \rho_2 n_2 \quad , \tag{7}$$

$$\frac{\partial n_1(r)}{\partial t} = \gamma_1(r)n_{FE} - [\nu_1 + \rho_1 + \Omega_1(r)]n_1(r) + \Omega_2(r)n_2(r) \quad , \tag{8}$$

$$\frac{\partial n_2(r)}{\partial t} = \gamma_2(r)n_{FE} - [\nu_2 + \rho_2 + \Omega_2(r)]n_2(r) + \Omega_1(r)n_1(r) \quad . \tag{9}$$

Here

g = the free exciton (FE) generation rate

n_{FE} = the FE density

ν_{FE} = the FE decay rate

n_i = the density of excitons bound to impurity atom i (i = 1.2); i.e., "1" = Aℓ(B), "2" = In

$n_i(r)$ = the density of excitons bound to impurity atom i with a nearest-neighbor impurity atom j ($\neq$ i) between r and r + dr

γ_i = the FE capture rate by impurity atom i

$\gamma_i(r)$ = the FE capture rate by impurity atom i with a nearest-neighbor impurity atom j ($\neq$ i) between r and r + dr

ν_i = the BE_i decay rate

ρ_i = the thermal release rate of BE_i

$\Omega_i(r)$ = the distance dependent exciton transfer rate from impuity atom i to impurity atom j ($\neq$ i) .

The FE capture rate $\gamma_i(r)$ is written

$$\gamma_i(r) = \sigma_i v_{th} [N_i(r) - n_i(r)] \quad , \tag{10}$$

where

σ_i = the FE capture cross section by impurity atom i

v_{th} = the FE thermal velocity .

Using Eq. (5), Eq. (10) in the absence of saturation [$n_i(r) \ll N_i(r)$] becomes

$$\gamma_i(r) = \sigma_i v_{th} N_i P_j(r) = \xi_i P_j(r) \quad , \quad (i \neq j) \tag{11}$$

where

$$\xi_i = \sigma_i v_{th} N_i \quad . \tag{12}$$

In this limit, we can obtain the steady state solution to Eqs. (7)-(9) by setting $dn_{FE}/dt = \partial n_i(r)/\partial t = 0$:

$$n_{FE} = g\{\nu_{FE} + \nu_1\xi_1[1 + \frac{\gamma_2}{\gamma_1} J_2 - J_1] + \nu_2\xi_2[1 + \frac{\gamma_1}{\gamma_2} J_1 - J_2]\}^{-1} \tag{13}$$

$$n_i(r) = n_{FE}\, \xi_i(r) \tag{14}$$

where

$$\xi_i(r) = \xi_i\{P_j(r) + [1 + \omega_1(r) + \omega_2(r)]^{-1} [\frac{\gamma_j}{\gamma_i} P_i(r)\omega_j(r) - P_j(r)\omega_i(r)]\} \tag{15}$$

with

$$\xi_i(r) = \frac{\gamma_i}{(\nu_i + \rho_i)} \sim N_i \tag{16}$$

$$\omega_i(r) = \frac{\Omega_i(r)}{(\nu_i + \rho_i)} \tag{17}$$

Integrating Eq. (14) over r, we obtain

$$n_i = \int_o^\infty n_i(r)dr$$

$$= \xi_i n_{FE} [1 + \frac{\gamma_j}{\gamma_i} J_j - J_i] \tag{18}$$

where

$$J_i = \int_o^\infty P_j(r)\omega_i(r) [1 + \omega_1(r) + \omega_2(r)]^{-1} \tag{19}$$

with

$$P_j(r) = 4\pi r^3 N_j \exp(-\frac{4\pi}{3} r^3 N_j) \tag{2}$$

C. <u>The Intensity Ratio of Bound Exciton Luminescence in Terms of Impurity Concentrations</u>

The ratio of PL intensity of BE_i to that of FE is given by

$$\frac{I_i}{I_{FE}} = \frac{f_i n_i}{f_{FE} n_{FE}} = \frac{f_i}{f_{EF}} \xi_i \ , \tag{20}$$

where f_i is the BE_i oscillator strength and f_{FE} is the FE oscillator strength. Taking the ratio I_1/I_2 with the help of Eqs. (18) and (20), we have

$$\frac{I_1}{I_2} = R_{12} \frac{N_1}{N_2} \ , \tag{1}$$

where

$$R_{12} = R_{12}^{o} \frac{1 + \eta^{-1} \cdot J_2 - J_1}{1 + \eta \cdot J_1 - J_2} \tag{21}$$

with

$$R_{12}^{o} = \frac{f_1}{f_2} \left[\frac{\xi_1 N_2}{\xi_2 N_1}\right] \tag{22}$$

= the value of R_{12} in the absence of the exciton transfer ,

and

$$\eta = \frac{\gamma_1}{\gamma_2} = \frac{\sigma_1 N_1}{\sigma_2 N_2} \quad . \tag{23}$$

We point out here that Eq. (21) contains the vital information which depends on the impurity concentrations and parameters such as the FE and BE oscillator strengths and the FE capture cross sections.

IV. THE EXCITON TRANSFER RATE

Before we compare this theory and the experimental results, we must calculate the exciton transfer rates $\Omega_i(r)$ in Eqs. (17),(19), and (21). The general topic of exciton transfer in solids has been considered extensively in the literature. Here we proceed with an original method used by Dexter[8] who considered excitation transfer through a modified Coulomb field from an excited impurity of one type (sensitizer) to another type of impurity (activator) in the ground state. In this case, the rate of energy transfer from an excited sensitizer to an activator is given by

$$W = \frac{2\pi}{\hbar} |\langle \psi_F | H_{int} | \psi_I \rangle|^2 \rho_F \tag{24}$$

where ψ_I is the wavefunction of the initial state in which the sensitizer is excited and the activator is in its ground state, ψ_F is the wavefunction of the final state in which the activator is excited and the sensitizer is in its ground state, H_{int} is the Coulomb interaction Hamiltonian, and ρ_F is the density of final states. Using the multipole expansion (in powers of r, the distance between the sensitizer and activator) of H_{int}, Dexter obtained the transfer rates for the dipole-dipole, dipole-quadrupole and exchange terms which results from the exchange integrals for properly antisymmetrized ψ_I and ψ_F . They are of the forms

$$\Omega_{dd}(r) = \Omega_{dd} r^{-6} \tag{25}$$

$$\Omega_{dq}(r) = \Omega_{dq} r^{-8} \tag{26}$$

$$\Omega_{ex}(r) = \Omega_{ex} e^{-r/\bar{a}} \qquad (\bar{a} = \text{the average } a_i, a_i') \quad . \tag{27}$$

In order to evaluate the quantities Ω_{dd}, Ω_{dq}, Ω_{ex} and $\bar{a}$, we need to have detailed knowledge of the wavefunctions for the Aℓ(B) BE and In BE. Pan, et al.[6], calculated the dissociation energy of an exciton bound to a neutral acceptor using a hydrogenic model wavefunction. This wavefunction was obtained by variation principle under the assumption that the exciton-neutral-acceptor complex consists of two relatively massive holes which are attracted to the negative acceptor center forming a "core" and a single relatively light electron repelled by the negative acceptor center. The hole-hole correlaion and correlation of the hole core with the electron were taken into account for calculating the correlation correction to the dissociation energy. In an excited sensitizer, two holes making up the "core" were represented by the 1s hydrogenic type wavefuntions

$$\phi_h(\vec{x}_{1h},\vec{x}_{2h}) = \frac{\alpha_h^3}{\pi} e^{-\alpha(r_{1h}+r_{2h})} , \tag{28}$$

where

$$r_{ih} = |\vec{x}_{ih}| \quad , \quad i = 1.2 \quad , \quad \alpha = \frac{1}{a} \quad .$$

The wavefunction for the electron excluding the spin states was given by

$$\phi'_e(\vec{x}) = [\frac{8\alpha'^3}{185\pi}]^{\frac{1}{2}}[1+(\alpha' r) + \frac{(\alpha' r)^2}{2}] e^{-\alpha' r} \tag{29}$$

where

$$r = |\vec{x}| \ .$$

The wavefunction for an activator in the ground state is given by

$$\phi(\vec{x}) = [\frac{\alpha^3}{\pi}]^{\frac{1}{2}} e^{-\alpha r} \tag{30}$$

In Eqs. (28)-(30), α, α' and α_h are the variational parameters.

Mitchard[5] pointed out that the exchange mechanism dominates the transfer process based on his measurements. We therefore attempted to calculate $\Omega_{ex}(r)$, leaving other terms for future investigations. It turned out that the matrix element $\langle\psi_F|H_{int}|\psi_I\rangle_{ex}$ can be calculated using the ellipsoidal coordinates and Neumann expansion.[9] Using the selection rules for the spin states and the wavefunction for the excited sensitizer given by Eq. (29) and that for the activator in the ground state given by Eq. (30), the exchange matrix element is written

$$\langle\psi_F|H_{int}|\psi_I\rangle = \int d\tau_{12}\phi_1^*(\vec{x}_2)\phi_2'^*(\vec{x}_1)\frac{e^2}{\kappa r_{12}}\phi_1'(\vec{x}_1)\phi_2(\vec{x}_2) \tag{31}$$

$$= \frac{2e^2}{185\kappa}[\alpha_1\alpha_2\alpha_1'\alpha_2']^{\frac{3}{2}} r^5 \sum_{\ell=0}^{4} \sum_{n_1,n_2=0}^{4} \omega_{n_1\ell}(\beta_1)\omega_{n_2\ell}(\beta_2) I^{\ell}_{n_1n_2}. \tag{32}$$

Here

κ = the dielectric constant of silicon

$$\left.\begin{aligned}
\omega_{0\ell}(\beta_i) &= -B_{2\ell}(\beta_i) + \frac{(\alpha_i'r)}{2} B_{3\ell}(\beta_i) - \frac{(\alpha_i'r)^2}{8} B_{4\ell}(\beta_i), && \ell = 0-4\\
\omega_{1\ell}(\beta_i) &= -\frac{(\alpha_i'r)}{2} B_{2\ell}(\beta_i) + \frac{(\alpha_i'r)^2}{4} B_{3\ell}(\beta_i), && \ell = 0-3\\
\omega_{2\ell}(\beta_i) &= B_{0\ell}(\beta_i) - \frac{(\alpha_i'r)^2}{2} B_{1\ell}(\beta_i), && \ell = 0-2\\
\omega_{3\ell}(\beta_i) &= \frac{(\alpha_i'r)}{2} B_{0\ell}(\beta_i) - \frac{(\alpha_i'r)^2}{4} B_{1\ell}(\beta_i), && \ell = 0,1\\
\omega_{40}(\beta_i) &= \frac{(\alpha_i'r)^2}{8} B_{00}(\beta_i)
\end{aligned}\right\} \tag{33}$$

$$\left.\begin{aligned}
B_{j\ell}(\beta) &= \left(\frac{d}{d\beta}\right)^j B_{0\ell}(\beta)\\
B_{0\ell}(\beta) &= \sqrt{(2\ell+1)\pi/\beta}\; I_{\ell+\frac{1}{2}}(\beta)
\end{aligned}\right\} \tag{34}$$

$I_{\ell+\frac{1}{2}}(\beta)$ = the modified Bessell function

and

$$I^{\ell}_{n_1n_2} = \int_1^{\infty} d\xi(\xi^2-1)^{-1} f_{n_1\ell}(\xi,\lambda_1) f_{n_2\ell}(\xi,\lambda_2) \tag{35}$$

with

$$f_{n,\ell}(\xi,\lambda) = \frac{1}{P_\ell(\xi)} \int_1^{\xi} dx\, P_\ell(x)\, e^{-\lambda x}\, x^n \tag{36}$$

and

$$\beta_i = \frac{r}{2}(\alpha_i' - \alpha_i), \qquad \lambda_i = \frac{r}{2}(\alpha_i + \alpha_i') \quad . \tag{37}$$

The simplest of $I^{\ell}_{n_1n_2}$ is I^{0}_{00} which is given by

$$I^{0}_{00} = \frac{1}{\lambda_1\lambda_2} e^{-(\lambda_1+\lambda_2)} [\varepsilon(\lambda_1) + \varepsilon(\lambda_2) - \varepsilon(\lambda_1+\lambda_2)], \tag{38}$$

where

$$2\varepsilon(x) = C + \ell n(2x) - e^{2x}Ei(-2x)$$

$$C = 0.577215 \ldots \quad \text{(Euler's constant)} \tag{39}$$

$$E_i(x) = \int_{-\infty}^{x} d\xi \, \frac{e^{\xi}}{\xi} \quad \text{for} \quad x < 0 \quad \text{(Exponential integral)}$$

The integral in Eq. (35) involves various exponential integrals which can be evaluated using several recurrence relations.[9] Equation (32) has some forty-two terms, more than one-half of them can be obtained using the recurrence relations.[9]

We note here that the exchange transfer rate $\Omega_{ex}(r)$ can also be obtained using a formula for tunneling of a particle in a potential well V(r) through a barrier in the WKB approximation in quantum mechanics.

$$\Omega_{ex}(r) = \frac{1}{\tau}\exp\left[-\frac{2}{\hbar}\left|\int_a^b p \, dr\right|\right] , \tag{40}$$

where τ is the period of classical oscillations within a potential well and

$$p = \left\{2m\left[E - V(r) - \frac{h^2(\ell+\frac{1}{2})^2}{2mr^2}\right]\right\}^{\frac{1}{2}} . \tag{41}$$

For a smoothly varying function p(r) , Eq. (40) becomes

$$\Omega(r) = \frac{1}{\tau} \exp\left[-\frac{2}{\hbar}\,\bar{p}(b-a)\right] , \tag{42}$$

where $\bar{p}$ is the average of p(r) over the range (b-a) . It can be shown that τ is related to the energy level of the exciton and the interval (b-a) is very nearly equal to r , the separation of two impurity atoms Aℓ(B) and In.

V. RECOMMENDATIONS

Equations (1), (21), and (22) contain vital information for interpreting the experimental results of Brown, et al. In the present work, we have used very simple wavefunctions for the exciton-neutral-acceptor complex in silicon. The BE transfer rates strongly depend on the BE wavefunctions. We make the following recommendations for future work:

1. Evaluate J_i in Eq. (21) in the present work since we were unable to do so because of time limits.
2. Obtain a more accurate BE wavefunction using an improved potential for the BE.

3. Recalculate the BE_i decay rate ν_i, the BE_i capture cross section σ_i, the BE_i oscillator strength f_i, the excition transfer rate $\Omega_i(r)$, and J_i using a more accurate BE wavefunction.
4. Since ν_i, σ_i and f_i can be measured from various experiments, one can justify the accuracy of the improved wavefunction if the results of these calculations agree with the experimental values.

REFERENCES

1. P. Blood and J.W. Orton, "The Electrical Characterization of Semiconductors," Rep. Prog. Phys., Vol. 41, pp. 157-257, 1978.

2. G.S. Mitchard and T.G. McGill, "Determination of Relative Impurity Concentrations Using Photoluminescence: A Case Study of the Si: (B,In) System," Appl. Phys. Lett., Vol. 37, pp. 959-961, 1980.

3. D.H. Brown, M.C. Ohmer, K.D. Beasley, J.J. Rome, G.J. Brown, and D.W. Fisher, "Optical and Electrical Characterization of Multiply-doped Silicon: A Study of the Si: (In,Aℓ) System," submitted to J. Appl. Phys. for publication.

4. V.O. Ziemelis and R.R. Parsons, "Sharp, Line Donor-Acceptor Pair Luminescence in Silicon," Can. J. Phys. Vol. 59, pp. 784-801, 1981.

5. G S. Mitchard, "Low Temperature Photoluminescence Properties of Silicon and Silicon-Germanium Alloys," Ph.D. thesis, California Institute of Technology, 1981 (unpublished).

6. D.S. Pan, D.L. Smith, and T.C. McGill, "Binding of an Exciton to a Neutral Acceptor," Solid State Commun., Vol. 18, pp. 1557-1560, 1976)

7. L.D. Landau and E.M. Lifshitz, Quantum Mechanics--Non-relativistic Theory (Pergamon Press, New York, 1958), pp. 171-178.

8. D.L. Dexter, "A Theory of Sensitized Luminescence in Solids," J. Chem. Phys., Vol. 53, pp. 836-850, 1953.

9. K. Rudenberg, "A Study of Two-Center Integrals Useful in Calculations on Molecular Structure. II. The Two-Center Exchange Integrals," J. Chem. Phys., Vol. 19, pp. 1459-1477, 1951.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

ERROR ANALYSIS OF THE KARHUNEN-LOEVE TRANSFORM IN VCG SIGNAL PROCESSING

Prepared by:	H. Troy Nagle, Jr., PhD, M.D.
Academic Rank:	Professor
Department and University:	Department of Electrical Engineering Auburn University
Research Location:	Air Force School of Aerospace Medicine, Clinical Sciences Division, Flight Medicine Branch
USAF Research Colleague:	LTC Gil D. Tolan, M.D.
Date:	August 24, 1982
Contract No:	F49620-82-C-0035

ACKNOWLEDGEMENT

The author would like to thank the Air Force Systems Command, AFOSR, and SCEEE for arranging this research opportunity. Special thanks are also sent to Dr. Gil Tolan who coordinated the work, to Dr. Ed Womble and Dr. Mel Lancaster of Systemics, Inc., for valuable discussions, and to Mr. Neil Keiser for help with the VAC 11/780.

ABSTRACT

by

H. Troy Nagle, Jr.

This paper derives error bounds for the computational accuracy of the Karhunen-Loeve Transform used in VCG signal processing. A procedure is proposed for selecting the number of transform coefficients, the wordlength of the coefficients, and the precision of the signal variables. The data compression ratio is discussed in terms of these parameters.

1.0 INTRODUCTION

During the last decade, the U.S. Air Force School of Aerospace Medicine has been conducting a research project in computer analysis of vectorcardiograms [1-6]. The analysis system employs various transformation and template algorithms for feature extraction, and fuzzy clustering for pattern recognition. The current system evolutionary state is displayed in Figure 1. It is noteworthy that the Karhunen-Loeve transformation is currently used in two portions of the system: QRS-T data compression and P-wave data compression. A summary of the KL transform is given below [1,7].

Let $\underline{x}$ be an N dimentional vector of time sequenced samples. We may represent $\underline{x}$ by

$$\underline{x} = \sum_{i=1}^{N} \alpha_i \underline{\phi}_i = \underset{(NXN)}{\Phi_N} \underset{(NX1)}{\underline{\alpha}_N} \qquad (1)$$

where α_i are the KL coefficients and $\underline{\phi}_i$ are orthonormal basis vectors. In scalar form equation (1) may be written

$$x(k) = \sum_{i=1}^{N} \alpha_i \phi_i(k)$$

the matrices and vectors in equation (1) are defined:

$$\Phi_N = \left[\underline{\phi}_1 \; \underline{\phi}_2 \cdots \underline{\phi}_N \right]$$

$$\underline{\alpha}_N = \left[\alpha_1 \; \alpha_2 \cdots \alpha_N \right]^T$$

$$\underline{x} = \left[x(1) \; x(2) \cdots x(N) \right]^T$$

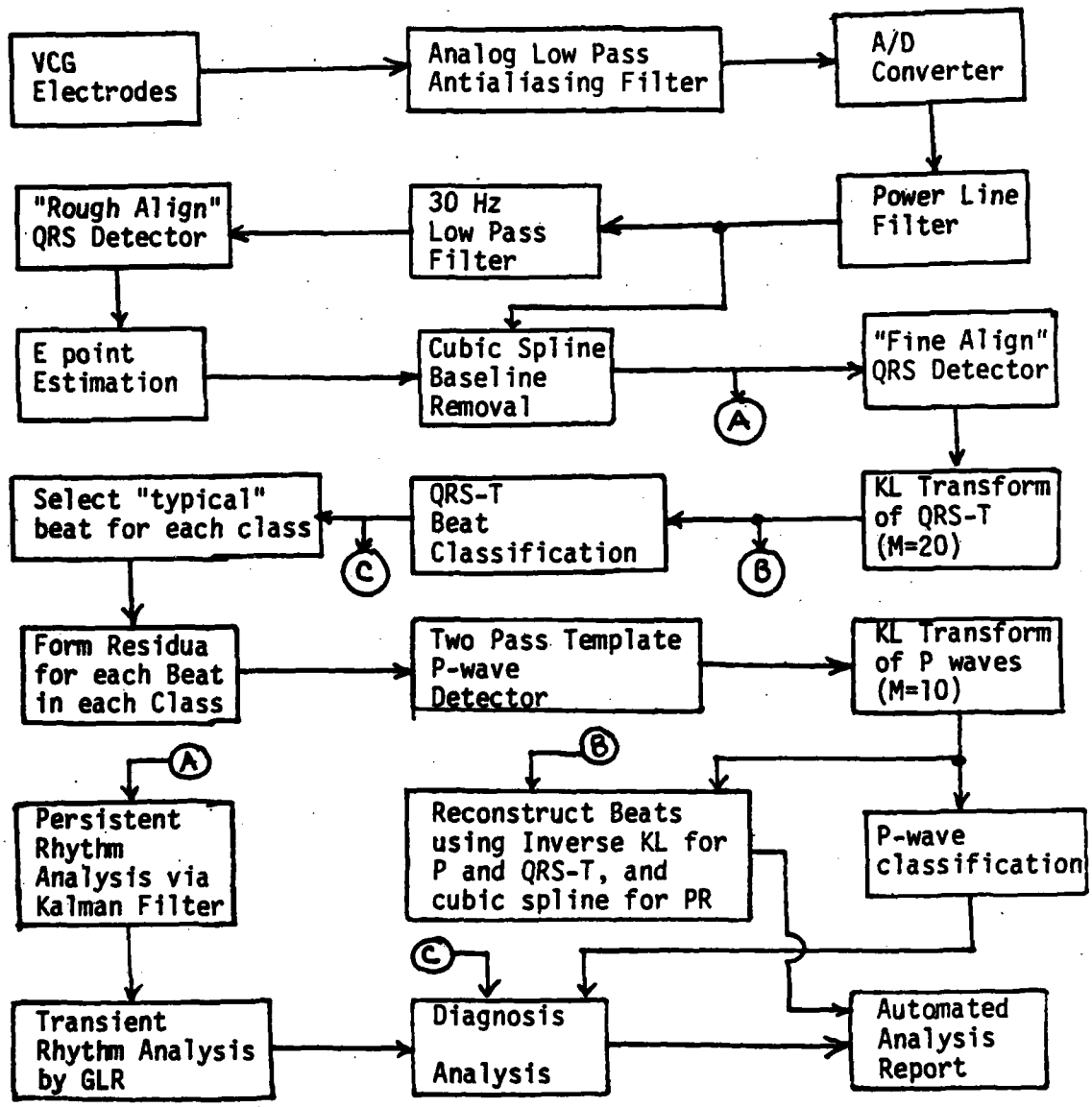


Figure 1. USAFSAM VCG Analysis System.

Since the functions $\underline{\phi}_i$ are orthonormal

$$\underline{\phi}_i^T \underline{\phi}_j = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases} \tag{2}$$

Consequently

$$\Phi_N^{-1} = \Phi_N^T$$

and we may calculate the KL coefficients

$$\alpha_i = \underline{\phi}_i^T \underline{x} = \sum_{k=1}^{N} x(k)\,\phi_i(k) \tag{3}$$

where $$\underline{\phi}_i^T = \left[\phi_i(1) \;\; \phi_i(2) \cdots \phi_i(N) \right]$$

In vector form

$$\underline{\alpha}_N = \underline{\Phi}_N^T \underline{x}$$

The basis vectors $\underline{\phi}_i$ are determined from the covarience matrix of $\underline{x}$, Σ_x

$$\Sigma_x \underline{\phi}_i = \lambda_i \underline{\phi}_i \tag{4}$$

The $\underline{\phi}_i$ and λ_i are the eigenvectors and eigenvalues of Σ_x .

The covariance matrix is calculated over a training set of K vectorcardiograms by

$$\Sigma_x = \frac{1}{K} \sum_{i=1}^{K} \underline{x}_i \underline{x}_i^T \tag{5}$$

The current KL transform of the QRS-T portion of the waveform uses a training set K=1000, and the number of time samples N=400.

In reality, equation (3) is not computed for 400 coefficients

Instead a greatly reduced number, M, may be used to represent the original signal $\underline{x}$ in the KL domain. Reference [7] has shown that if M coefficients are used, the minimum mean squared error

$$\bar{\varepsilon}^2(M) = E\left\{ \| \underline{x} - \underline{\hat{x}}_M \| \right\} = \sum_{i=M+1}^{N} \lambda_i \tag{6}$$

where

$$\underline{\hat{x}}_M = \sum_{i=1}^{M} \alpha_i \underline{\phi}_i + \sum_{i=M+1}^{N} \underline{\phi}_i^T E\{\underline{x}\} \underline{\phi}_i \tag{7}$$

If we subtract the expected value $E\{\underline{x}\}$ from the input waveform $\underline{x}$ before taking the KL transform,

$$\underline{\hat{x}}_M = \sum_{i=1}^{M} \alpha_i \underline{\phi}_i = \underset{(N \times M)}{\Phi_M} \underset{(M \times 1)}{\underline{\alpha}_M} \tag{8}$$

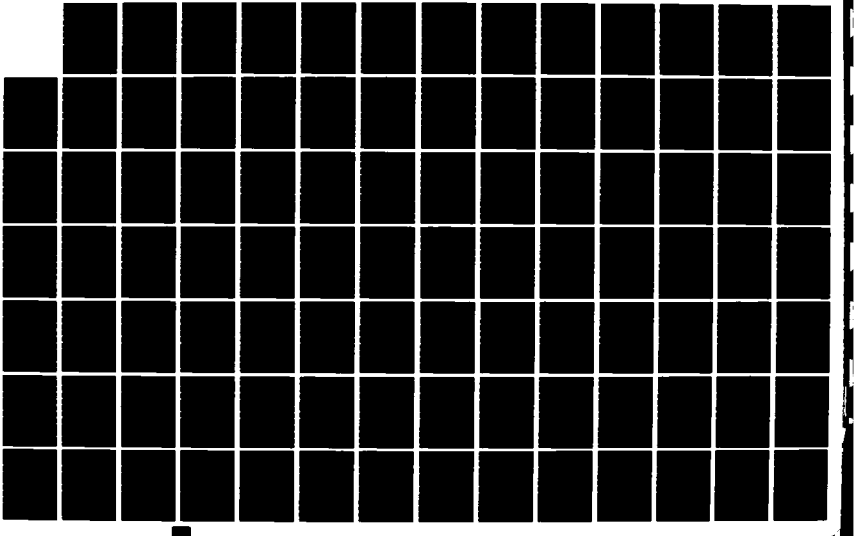
Equation (8) tells us that for a given value of M, the mean square error in reconstructing $\underline{x}$ (using only M terms in the reconstruction) is the sum of the eigenvalues λ_i for $N \geq i > M$.

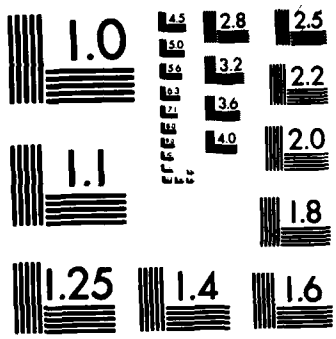
Note that the eigenvalues are monotonically decreasing with increasing i.

AD-A130 770 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM (1982) RESEARCH REPORTS VOLUME..(U) SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION INC S..

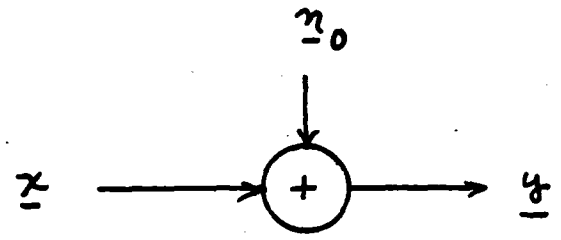
4/11

UNCLASSIFIED W D PEELE ET AL. OCT 82 AFOSR-TR-83-0614 F/G 5/1 NL

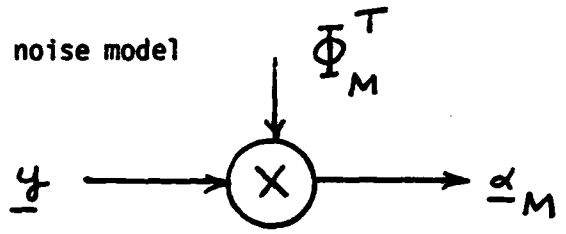




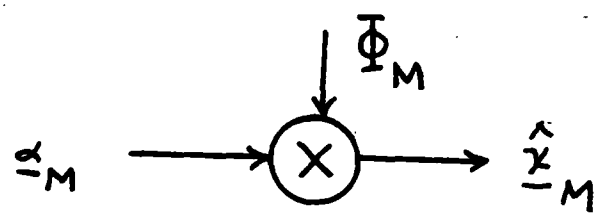
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



a) noise model



b) The KL Transform



c) Reconstructed Waveform

Figure 2.

2.0 ERROR MODEL

A question frequently raised in ECG/VCG analysis is the data compression and accuracy of the KL transform. Equation (6) gives us the minimum mean squared error caused by truncating the series of equation (1) from N to M terms. In the USAFSAM project, N is typically 400 and M is 20, so that the mean squared error is

$$\bar{\varepsilon}^2(M) = \bar{\varepsilon}^2(20) = \sum_{i=21}^{400} \lambda_i \tag{9}$$

It has been estimated that 98% of the signal energy is represented by the first 20 eigenvalues [1], or

$$\frac{\bar{\varepsilon}^2(20)}{\bar{\varepsilon}^2(0)} \doteq .02 \tag{10}$$

The series truncation error mentioned above is but one of many error sources in the ECG/VCG analysis system. Consider the simplified model of Figure 2. Let $\underline{x}$ be the pure bioelectric potential generated by the heart muscle. Before this potential can analyzed it is corrupted in varying degrees by

(1) power line interference
(2) respiratory changes in the baseline
(3) muscle contraction noise
(4) changes in body position
(5) electrode contact noise
(6) electrode placement artifacts
(7) lead wire motion (coupling)
(8) instrumentation electronic noise, and
(9) quantization error in A/D conversion.

Each of these noise sources has its own characteristics. In Figure 2a, all of these effects have been lumped together and modeled as an additive noise source, $\underline{n}_0$:

$$\underline{y} = \underline{x} + \underline{n}_0 \tag{11}$$

The measured signal is $\underline{y}$ which is available as a sequence of N samples. The goal of the KL transform in Figure 2b and c is to estimate the true biopotential $\underline{x}$ from the noisy measurement $\underline{y}$.

First let us examine the KL transform

$$\underline{\alpha}_M = \underline{\Phi}_M^T \underline{y} \tag{12}$$

The details of this calculation are presented in Figure 3. Notice that multiple additive noise sources are modeled. Each of these error sources will now be identified:

(1) roundoff error in representing $\phi_i(k)$

$$\left[\phi_i(k)\right]_c = \phi_i(k) + n_{1i}(k)$$

where $\left[\phi_i(k)\right]_c$ is the finite wordlength representation of $\phi_i(k)$.

(2) roundoff error $n_{2i}(k)$ in representing the product of $\left[\phi_i(k)\right]_c$ times $\left[y(k)\right]_c$.

(3) roundoff error n_{3i} in representing the sum of product terms, and

(4) roundoff error n_{4i} in storing the minimum wordlength $\left[\alpha_i\right]_c$ in the data base.

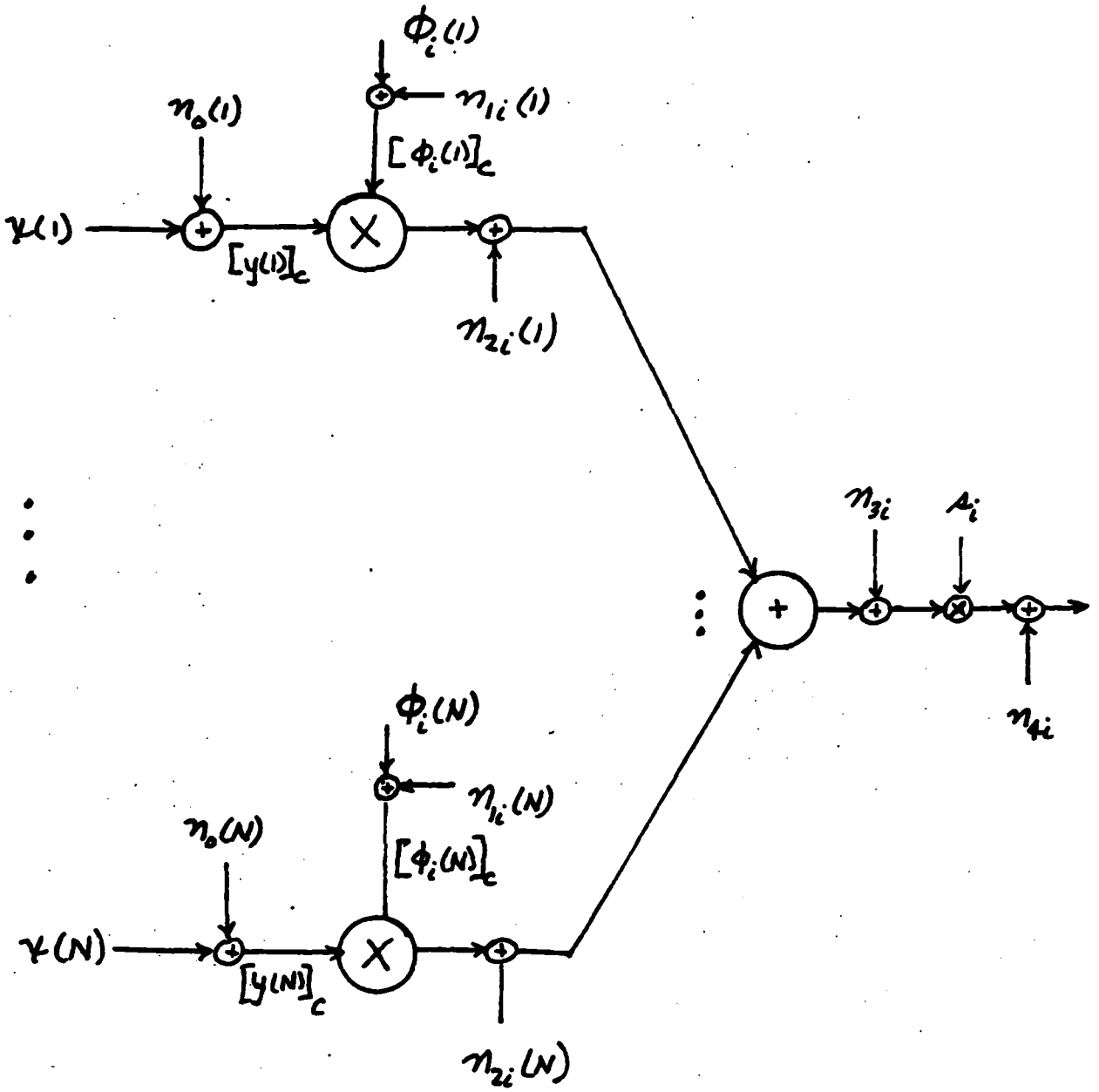


Figure 3. KL transform, i=1 to M.

At this point we should emphasize that a scaling factor, A_i , has been included just prior to noise source n_{4i} . The purpose of this scaling factor is to ensure that

$$|[\alpha_i]_c| \leq 1$$

In analyzing the numerical computations of Figure 3 we will adopt the convenient convention that the magnitude of all signals and coefficients will be scaled so that their magnitudes are less than or equal to unity,

From Figure 3 we may write for i = 1 to M:

$$[\alpha_i]_c = A_i \left\{ \sum_{k=1}^{N} \left[(x(k) + n_0(k)) \left(\phi_i(k) + n_{1i}(k) \right) + n_{2i}(k) \right] + n_{3i} \right\} + n_{4i}$$

$$\frac{[\alpha_i]_c}{A_i} = \sum_{k=1}^{N} x(k)\,\phi_i(k) + \sum_{k=1}^{N} \left[x(k)\,n_{1i}(k) + \phi_i(k)\,n_0(k) + n_0(k)\,n_{1i}(k) + n_{2i}(k) \right] + n_{3i} + \frac{n_{4i}}{A_i}$$

Using equation (3)

$$\frac{[\alpha_i]_c}{a_i} = \alpha_i + \epsilon_{\alpha_i} \tag{13}$$

where

$$\epsilon_{\alpha_i} = \sum_{k=1}^{N} \left[x(k)\eta_{1i}(k) + \phi_i(k)\,\eta_0(k) + \eta_0(k)\eta_{1i}(k) + \eta_{2i}(k) \right] + \eta_{3i} + \frac{\eta_{4i}}{a_i} \tag{14}$$

Consequently, the error in representing a KL coefficient is ϵ_{α_i} which is a function of the noise in the measured signal, η_0 , the noise in calculating the orthonormal functions, η_{1i} , the noise in rounding product terms, η_{2i} , the noise in rounding the sum of product terms, η_{3i} , and the noise in further rounding the coefficient for storage in the data base, η_{4i} . It should be noted that η_{3i} and η_{4i} are both rounding operations and that one will dominate and exclude the other from the relation of equation (14).

Next, let us examine the inverse KL transform of Figure 2

$$\hat{\underline{x}}_M = \Phi_M \, \underline{\alpha}_M \tag{15}$$

We may model this calculation as shown in Figure 4. From the figure, for k = 1 to N, we may write

$$\hat{x}_M(k) = r_k \left\{ \sum_{i=1}^{M} \left[a_i(\alpha_i + \epsilon_{\alpha_i})(\phi_i(k) + \eta_{1i}(k)) + \eta_{5i}(k) \right] + \eta_6(k) \right\} + \eta_7(k)$$

where r_k is a scaling factor.

$$\frac{\hat{x}_M(k)}{r_k} = \sum_{i=1}^{M} a_i \alpha_i \phi_i(k) + \sum_{i=1}^{M} \left[a_i \alpha_i \eta_{1i}(k) + a_i \epsilon_{\alpha_i} \phi_i(k) + a_i \epsilon_{\alpha_i} \eta_{1i}(k) + \eta_{5i}(k) \right] + \eta_6(k) + \frac{\eta_7(k)}{r_k} \tag{16}$$

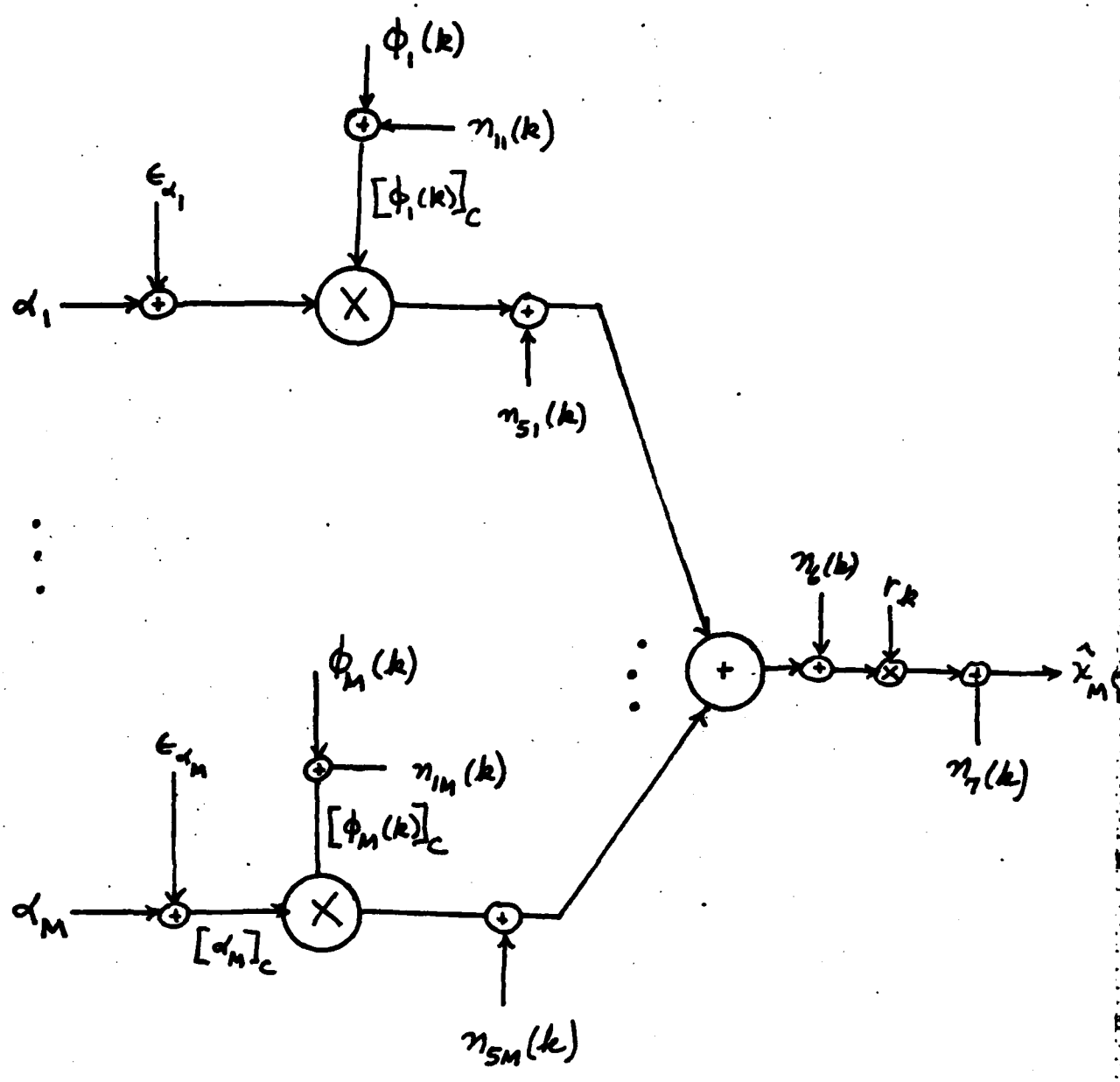


Figure 4: Inverse KL Transform, k = 1 to N.

But from equation (1)

$$x(k) = \sum_{i=1}^{N} \alpha_i \phi_i(k)$$

So

$$\sum_{i=1}^{M} \alpha_i \phi_i(k) = x(k) - \sum_{i=M+1}^{N} \alpha_i \phi_i(k) \tag{17}$$

and if

$$a_i = a' \geq \max(\alpha_1, \alpha_2, \ldots, \alpha_M)$$

$$\frac{\hat{x}_M(k)}{r_k} = a' x(k) - a' \sum_{i=M+1}^{N} \alpha_i \phi_i(k)$$

$$+ a' \left\{ \sum_{i=1}^{M} \alpha_i n_{1i}(k) + \epsilon_{\alpha_i} \phi_i(k) + \epsilon_{\alpha_i} n_{1i}(k) \right\}$$

$$+ \sum_{i=1}^{M} n_{5i}(k) + n_6(k) + \frac{n_7(k)}{r_k}$$

Therefore, the total error model is

$$\frac{\hat{x}(k)}{r_k a'} = x(k) + \epsilon_x(k) \tag{18}$$

where

$$\epsilon_x(k) = \sum_{i=1}^{M} \left[\alpha_i n_{1i}(k) + \epsilon_{\alpha_i} \phi_i(k) + \epsilon_{\alpha_i} n_{1i}(k) + \frac{n_{5i}(k)}{a'} \right]$$

$$+ \frac{n_6(k)}{a'} + \frac{n_7(k)}{r_k a'} - \sum_{i=M+1}^{N} \alpha_i \phi_i(k) \tag{19}$$

Again let us emphasize that the inverse KL transform is using equation (8) instead of equation (7), which assumes that the expected value of $\underline{x}$ has been subtracted from the measured input $\underline{y}$ before any computations are made. We may add this vector back to $\hat{\underline{x}}_M$ to complete the computations.

The scaling factor r_k in equation (18) is chosen to normalize $\hat{x}(k)$

$$|\hat{x}(k)| \leq 1 \tag{20}$$

But the input sequence $x(k)$ is assumed to be normalized, so that if we choose

$$r_k = \frac{1}{s'} \tag{21}$$

for all k

$$\hat{x}(k) = x(k) + \epsilon_x(k) \tag{22}$$

and

$$\epsilon_x(k) = \sum_{i=1}^{M} \left[\alpha_i n_{1i}(k) + \epsilon_{\alpha_i} \phi_i(k) + \epsilon_{\alpha_i} n_{1i}(k) + \frac{n_{5i}(k)}{s'} \right] + \frac{n_6(k)}{s'} + n_7(k) - \sum_{i=M+1}^{N} \alpha_i \phi_i(k) \tag{23}$$

3.0 ERROR BOUNDS [8]

The computations of the transfroms in Figures 3 and 4 may be carried out in various forms of arithmetic. Here we consider the fixed-point, 2's complement number system used on most commercially available computers today. Reference [8] has shown that a properly biased, successive-approximation A/D converter of b_1+1 bits (b_1 magnitude bits plus a sign bit) performs as a roundoff quantizer. Hence, if the input to the A/D is scaled so that its magnitude is always less than or equal to unity, the mean squared error introduced by the A/D ($n_{A/D}$) is approximated by

$$\sigma^2_{n_{A/D}} \doteq 2^{-2b_1}/12 \tag{24}$$

Furthermore, reference [8] has shown that if the product of a two's complement coefficient of a+1 bits and a two's complement variable of b_2+1 bits is rounded to b_2+1 bits, the error introduced is characterized by

$$\sigma^2_{e_r} = E\{e_r^2\} - E^2\{e_r\}$$
$$= (2^{-2b_2}/12)(1 - 2^{-2a})$$

and since in practice a is always greater than 4 , $2^{-2a} \ll 1$ and

$$\sigma^2_{e_r} \doteq 2^{-2b_2}/12 \tag{25}$$

Equation (25) tells us that if the product which is represented by $a+b_2+1$ bits is rounded to b_2+1 bits, the mean squared error introduced is $2^{-2b_2}/12$, the same format found for the A/D converter. We may also state that the magnitude of this roundoff error is

$$2^{-b_2}/2 \geq e_r > -2^{-b_2}/2 \tag{26}$$

3.1 Error Bounds for the KL Transform

Equation (14) models the error in the KL transform. Let us consider the relative magnitude of each term in the expression:

(1) $x(k)$, normalized variables of magnitude

$$|x(k)| \leq 1 \tag{27}$$

(2) $\phi_i(k)$, orthonormal functions defined by equation (4). From equation (2)

$$\sum_{k=1}^{N} \phi_i^2(k) = 1 \tag{28}$$

so that

$$|\phi_i(k)| \leq 1 \tag{29}$$

(3) $n_0(k)$, input noise from many sources; in the noiseless environment the A/D will still introduce error bounded by equation (22)

$$|n_0(k)| \leq 2^{-b_1}/2 \tag{30}$$

$$\sigma_{n_0}^2 = 2^{-2b_1}/12$$

(4) s_i , the scaling factor

$$s_i = s' \geq \max(\alpha_1, \alpha_2, \ldots, \alpha_M)$$

we can restrict s' to be a power of 2 so that

$$s_i = s' = 2^{-c} \tag{31}$$

and

$$|\alpha_i| \leq 2^{-c} \tag{32}$$

(5) $n_{1i}(k)$, error in representing $\phi_i(k)$ by

$[\phi_i(k)]_c$. If b_2+1

bits are used, and high precision is used to compute off-line by equation (4), the error is bounded by equations (25) and (26).

(6) $n_{2i}(k)$, roundoff of the product terms to b_2+1 bits is also bounded by equation (22). If double length products ($a+b_2+1$ bits) are used in the summation, then this error source may be eliminated!

(7) n_{3i}, roundoff of the sum to b_2+1 bits. n_{4i} is additional roundoff of the scaled KL coefficient to a bits. Only one noise source will actually be inplemented

$$2^{-a}/2 \geq n_{4i} > -2^{-a}/2 \tag{33}$$

where

$$a \leq b_2$$

To achieve the greatest accuracy we recommend using double-length products in sum, and hence equation (14) becomes

$$\epsilon_{4i} = \sum_{k=1}^{N} \left[\chi(k)\, n_{1i}(k) + \phi_i(k)\, n_0(k) + n_0(k)\, n_{1i}(k)\right] + \frac{n_{4i}}{a'} \tag{34}$$

But by equations (27) and (29)

$$|\epsilon_{d_i}| \le \sum_{k=1}^{N} \left[|n_{1_i}(k)| + |n_0(k)| + |n_0(k)||n_{1_i}(k)| \right] + \frac{|n_{d_i}|}{s'} \tag{35}$$

But $|n_{1_i}(k)| << 1$, so

$$|\epsilon_{d_i}| \le \sum_{k=1}^{N} \left[|n_{1_i}(k)| + |n_0(k)| \right] + \frac{|n_{d_i}|}{s'}$$

$$\le N \left[|n_{1_i}(k)|_{max} + |n_0(k)|_{max} \right] + \frac{|n_{d_i}|_{max}}{s'} \tag{36}$$

Hence

$$|\epsilon_{d_i}| \le N |n_0(k)|_{max} + N(2^{-b_2}/2) + \frac{(2^{-a}/2)}{2^{-c}} \tag{37}$$

If we represent the input noise as being generated solely by the A/D (the noiseless case!);

$$|\epsilon_{d_i}| \le N(2^{-b_1} + 2^{-b_2})/2 + 2^{c-a}/2$$

and if $b = b_1 = b_2$,

$$|\epsilon_{d_i}| \le 2^{-b} N + 2^{c-a-1} \tag{38}$$

This relation suggests that to balance the error contribution of rounding variables and coefficients, that

$$2^{c-a} = 2^{-b+1} N \tag{39}$$

Example

Suppose we want to limit the error in representing $[\alpha_i]_c$ to less than 1 part in 1000.

$$[\alpha_i]_c = s' (\alpha_i + \epsilon_{d_i}) \tag{40}$$

So that

$$[\alpha_i]_c - \Delta' \alpha_i = \Delta' \epsilon_{\alpha_i}$$

and

$$\Delta' |\epsilon_{\alpha_i}| \leq 1/1000$$

Consequently, we may set

$$\Delta' (2^{-b_N} + 2^{c-a-1}) = 10^{-3} \tag{41}$$

But $\Delta' = 2^{-c}$ and $10^{-3} \doteq 2^{-10}$

so for balanced contributions of each term in equation (41)

$$2^{-a-1} \leq 2^{-10}/2$$

and

$$a = 10$$

In a similar manner

$$2^{-(b+c)} N \leq 2^{-10}/2$$

But in our application N = 400

so

$$2^{-(b+c)} \leq 2^{-10}/800$$

so

$$b+c = 20$$

Recall that $|\alpha_i|_{max} \leq 2^{-c}$ which must be computed from a typical waveform from the training set. We should also emphasize that the upper error bound of Equation (38) is very pessimistic.

A more practical approach to analyzing the error in the KL transform is to employ a statistical approach to derive a bound on the mean-squared error. Let us approximate equation (34) by

$$\epsilon_{\alpha_i} \doteq \sum_{k=1}^{N} \left[\chi(k)\, n_{1i}(k) + \phi_i(k)\, n_0(k) \right] + \frac{n_{\alpha_i}}{\Delta'} \tag{42}$$

From reference [8]

$$\sigma^2_{\epsilon_{\alpha_i}} \doteq \sum_{k=1}^{N} \left[\chi^2(k)\, \sigma^2_{n_{1i}} + \phi_i^2(k)\, \sigma^2_{n_0} \right] + \frac{\sigma^2_{n_{\alpha_i}}}{(\Delta')^2} \tag{43}$$

$$\sigma_{\epsilon_{d_i}}^2 \leq N\sigma_{n_{1i}}^2 + \sigma_{n_0}^2 + \frac{\sigma_{n_{d_i}}^2}{(s')^2} \tag{44}$$

Substitutiong known values

$$\sigma_{\epsilon_{d_i}}^2 \leq \sigma_{n_0}^2 + N(2^{-2b_2}/12) + 2^{2(c-a)}/12 \tag{45}$$

Again assuming the case of a noiseless VCG,

$$\sigma_{n_0}^2 = 2^{-2b_1}/12$$

and

$$\sigma_{\epsilon_{d_i}}^2 \leq \frac{1}{12}\left[2^{-2b_1} + N\,2^{-2b_2} + 2^{2(c-a)}\right] \tag{46}$$

Once again we may balance the effect of a, b_1, and b_2 by setting

$$2^{-b_1} \doteq 2^{-2b_2}(N) \doteq 2^{2(c-a)} \doteq 4\,\sigma_{\epsilon_{d_i}}^2 \tag{47}$$

Example

Suppose we want to bound the error in representing $[\alpha_i]_c$ to as in the previous example. Then

$$(s')^2\sigma_{\epsilon_{d_i}}^2 \doteq 2^{-20}/12$$

and by equation (47)

$$2^{2(c-a)} \leq 4(2^{-20}/12)\,2^{2c} \tag{48}$$

or

$$2^{-2a} \leq 2^{-21}(4/3)$$

and

$$a = 11$$

The second expression in equation (47) becomes

$$2^{-2b_2}(N) \leq 4(2^{-20}/12)\,2^{2c} \tag{49}$$

and with N = 400,

$$2^{-2(b_2+c)} \leq 2^{-20}/1200$$

and

$$b_2 + c = 16.$$

Note that the variable wordlength, b_2, is somewhat shorter (4 bits to be exact) than in the last example.

The effect of the A/D converter, b_1, is

$$2^{-b_1} \leq 4\,(2^{-20}/12)\,2^{2C}$$

$$2^{-2(b_1+C)} \leq (4/3)\,2^{-22}$$

Thus,

$$b_1 + C = 11$$

3.2 Error Bounds for the Inverse KL Transform

Now let us repeat the analysis of the last section on Figure 4. Equation (23) may be approximated, using the same assumptions of the last section, by

$$\epsilon_x(k) \doteq \sum_{i=1}^{M} \left[\alpha_i n_{1i}(k) + \phi_i(k)\epsilon_{\alpha_i}\right] + n_x(k) - \sum_{i=M+1}^{N} \alpha_i \phi_i(k) \tag{50}$$

upper bound error analysis yields

$$|\epsilon_x(k)| \leq \sum_{i=1}^{M} |\alpha_i|\,|n_{1i}(k)| + \sum_{i=1}^{M} |\phi_i(k)|\,|\epsilon_{\alpha_i}| + |n_x(k)| + \sum_{i=M+1}^{N} |\alpha_i|\,|\phi_i(k)|$$

But using the maximum values

$$|\epsilon_x(k)| \le M\left[2^{-b_2-c-1}\right] + M\left[2^{-a+c-1}\right] + 2^{-b_2}/2 + 2^c(N-M) \tag{51}$$

The last term in (51) gives a very large upper bound and makes this relationship useless in the practical sense. However, mean-squared error analysis yields

$$\sigma_{\epsilon_x}^2 = \sum_{i=1}^{M}\left[\alpha_i^2\sigma_{n_1}^2 + \phi_i^2(k)\sigma_{\epsilon_{d_i}}^2\right] + \sigma_{n_2}^2 + \bar{\epsilon}^2(M) \tag{52}$$

where $\bar{\epsilon}^2(M)$ is the truncation error of equation (6).
Consequently,

$$\sigma_{\epsilon_x}^2 \le \sigma_{n_1}^2\left(\sum_{i=1}^{M}\alpha_i^2\right) + \sigma_{\epsilon_{d_i}}^2\left(\sum_{i=1}^{M}\phi_i^2(k)\right) + \sigma_{n_2}^2 + \bar{\epsilon}^2(M) \tag{53}$$

Since $\sum_{k=1}^{N}\phi_i^2(k) = 1$, we may approximate $\sum_{i=1}^{M}\phi_i^2(k) < 1$, and

$$\sigma_{\epsilon_x}^2 \le \frac{2^{-2b_2}}{12}\left(M2^{-2c}\right) + \left(\frac{2^{-2b_1} + N2^{-2b_2} + 2^{2(c-a)}}{12}\right)(1) + \left(\frac{2^{-2b_2}}{12}\right) + \sum_{i=M+1}^{N}\lambda_i \tag{54}$$

$$\le \frac{2^{-2b_2}}{12}\left(M2^{-2c} + N + 1\right) + \frac{2^{-2b_1}}{12} + \frac{2^{-2a}}{12}\left(2^{2c}\right) + \sum_{i=M+1}^{N}\lambda_i \tag{55}$$

4.0 GUIDELINES FOR KL IMPLEMENTATION

In order to employ equation (56) to define the parameters of the KL transform, we propose that each of the four terms have approximately the same impact on the expected squared error. Therefore, the following algorithm is offered.

Algorithm

1. Estimate the value of c from the training set
   $2^{-c} \geq \max(\alpha_1, \alpha_2, \ldots, \alpha_M)$
   for a large value of M, perhaps M = N.
2. From clinically significant waveform changes, estimate an acceptable $\sigma^2_{\epsilon_x}$ for the noiseless VCG.
3. Use Term_1 to calculate the value of M such that
   $$\sum_{i=M+1}^{N} \lambda_i \leq \sigma^2_{\epsilon_x}/4 \qquad (57)$$
4. Use Term_2 to calculate the wordlength of the A/D converter, b_1+1
   $$2^{-b_1} \leq 3\sigma^2_{\epsilon_x} \qquad (58)$$
5. Use Term_3 to calculate the KL coefficient wordlength, a + 1
   $$2^{-2a} \leq 2^{-2c} 3\sigma^2_{\epsilon_x} \qquad (59)$$

6. Use Term_4 to calculate the computer wordlength, b_2+1

$$2^{-b_2} \leq 3\sigma_{\epsilon_x}^2 / (M2^{-2c}+N+1) \tag{60}$$

7. Calculate the KL data compression ratio

$$DCR = \frac{(N)(b_1+1)}{(M)(a+1)} \tag{61}$$

Example

Let us apply the algorithm to the analysis of a VCG. In particular the KL transform and waveform reconstruction of the USAFSAM project is an appropriate example.

Reference [9] gives us some details for a KL expansion of N = 200 samples with orthonormal functions calculated from a training set of 936 VCG's. A table of the computed eigenvalues follows:

i	λ_i	i	λ_i	i	λ_i	i	λ_i
1	.36152	6	.04939	11	.00659	16	.00212
2	.26071	7	.03200	12	.00537	17	.00170
3	.12525	8	.02663	13	.00358	18	.00129
4	.11359	9	.00983	14	.00307	19	.00123
5	.07243	10	.00764	15	.00266	20	.00080

Another result of reference [9] is that the truncation error in the KL expansion with M = 60 is about .1%. Now let us apply the algorithm.

1. Calculate the scaling parameter, c.

$$\lambda_i = E\{\alpha_i^2\}$$

but λ_1, is the largest eigenvalue.

From the table above

$$\lambda_1 = .36152$$

or

$$\sqrt{E\{\alpha_i^2\}} \doteq .6$$

so we may set

$$2^{-c} = 1 > \max(\alpha_1, \alpha_2, \ldots, \alpha_M)$$

so

$$c = 0.$$

2. The American Heart Association has estimated that a clinically significant change in an ECG results if less than 9 bits is used in its representation. So let us adopt $\sigma_{\epsilon_x}^2 = 2^{-18}$

3. To calculate M we evaluate

$$\sum_{i=M+1}^{N} \lambda_i \leq 2^{-18}/4 \doteq 10^{-6}$$

From reference [9]

$$\sum_{i=61}^{200} \lambda_i \doteq 10^{-6}$$

So

$$M = 60.$$

4. The A/D wordlength

$$2^{-2b_1} \leq 3 \times 2^{-18}$$

So

$$b_1 = 9$$

and a 10 bit A/D converter is needed.

5. The KL coefficient wordlength

$$2^{-2a} \leq 3(2^{-2c})(2^{-18}) = 3 \times 2^{-18}$$

So

$$a = 9$$

Thus 10 bit KL coefficients are needed.

6. The computer wordlength

$$2^{-b_2} \leq 3(2^{-18})/(M+N+1)$$

$$\leq \left(\frac{3}{261}\right)2^{-18} = \left(\frac{128}{87}\right)2^{-25}$$

so

$$b_2 = 13$$

and a 14 bit computer is needed.

7. The data compression ratio is

$$DCR = \frac{(N)(b_1+1)}{(M)(a+1)}$$

$$= \frac{(200)(10)}{(60)(10)}$$

$$= 3.33$$

5.0 CONCLUSIONS

An algorithm for determining the parameters of the KL transform has been derived. The algorithm fits the data from reference [9] very well. Additional studies to verify the algorithm are suggested. The algorithm suggests that a KL transform using 60 terms is needed with 10 bit KL coefficients to preserve the accuracy of a sequence of 200 samples from a 10 bit A/D converter.

6.0 RECOMMENDATIONS

The USAFSAM will conduct a field trial of the VITAL system in 1983. The VITAL system will be compared to commercially available measurement analysis systems. Problems that must be solved before VITAL can compete in the trial are (See Figure 1):

1) Accurate documentation of the current system must be obtained from S^2I.

2) The subunits for the overall algorithm must be verified

   a) KL of QRS-T
   b) P-wave detection
   c) Classification algorithms

3) A measurements routine must be written, or adapted, for providing inputs to the Diagnosis section.

4) A Diagnosis section must be designed, coded and tested.

Problems 3 and 4 are major ones and could delay a successful field trial until 1984 or 1985.

REFERENCES

[1] M.E. Womble, et al., "Data Compression for Storing and Transmitting ECG's/VCG's", Proc. IEEE, Vol. 65, No. 5, May 1977, pp. 702 - 706.

[2] D.E. Gustafson, et al., "A Statistical Approach to Rhythm Diagnosis of Cardiograms", Proc. IEEE, vol. 65, no. 5, May 1977, pp. 802 - 804.

[3] C.R. Meyer and H.N. Keiser, "Electrocardiogram Baseline Noise Estimation and Removal Using Cubic Splines and State-Space Computation Techniques", Comp. Bio. Res., vol. 10, 1977, pp. 459 - 470.

[4] D.E. Gustafson, et al., "ECG/VCG Rhythm Diagnosis Using Statistical Signal Analysis, Parts I and II", IEEE Trans. BME, vol. BME - 25, No. 4., July 1978, pp. 344 - 361.

[5] G.D. Tolan, et al., "Space Technology Applied to VCG Analysis", Proc. of Computers in Cardiology, IEEE no. CH 1606 - 3, 1980, pp. 395 - 397.

[6] G.D. Tolan, "Vectorcardiogram Serial Comparison Using Modern Communication and Control Theory Techniques", Proc. of Computers in Cardiology, IEEE No. 0276 - 6574, 1981, pp. 493 - 496.

[7] K. Fukunaga, Introduction to Statistical Pattern Recognition. New York: Academic Press, 1972.

[8] C.L. Phillips and H. T. Nagle, Jr., Analysis and Design of Digital Control Systems. Englewood Cliffs, NJ: Prentice - Hall, in production.

[9] D. Gustafson, et al., "Investication of Signal Analysis Techniques for ECG/VCG Classification, vol. I, Analytical and Numerical Studies, "Final Report, Contract F33615 - 77 - C - 0606, Scientific Systems, Inc., Cambridge, Mass., Jan. 1978.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

SIMULATION OF ADAPTIVE NETWORKS BASED ON REST PRINCIPLE AND HETEROSTATIC MODELS

Prepared by:	Eugene M. Norris
Academic Rank:	Associate Professor of Computer Science
Department and University	Mathematical Sciences Department George Mason University
Research Location:	Information Processing Technology Branch System Avionics Division Avionics Laboratory Air Force Wright Aeronautical Laboratories Wright-Patterson AFB, Ohio 45433
USAF Research Colleague:	Dr. A. Harry Klopf
Date:	29 July 1982
Contract No:	F49620-82-C-0035

SIMULATION OF ADAPTIVE NETWORKS BASED ON REST PRINCIPLE AND HETEROSTATIC MODELS

by

Eugene M. Norris

ABSTRACT

The objective of the work reported on here is the establishment of a preliminary estimate of the relative capabilities of networks of adaptive components in the performance of closed-loop learning tasks. A simulation test bed was written in 1966 FORTRAN and a scheme for uniformly describing adaptation algorithms based on recent work of Sinclair and of Klopf was devised. Results of a one-dimensional landmark learning experiment are discussed. It is discovered that the relation between an adaptive network and its environment is subtle and that much work is likely needed before a workable set of design criteria for networks and environments can be developed. The effects of introducing inhibition are likewise in need of further study.

ACKNOWLEDGMENT

The author expresses his thanks to Air Force Systems Command, Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for the opportunity to pursue this research at the Avionics Laboratory, A particular expression of gratitude is made to Dr. Don Moon, Head, Information Processing Technology Branch, for his hospitality in providing office and laboratory facilities.

The guidance provided by Dr. A. Harry Klopf has been a source of great stimulation and pleasure. It has been an honor to collaborate with him on this work.

I. INTRODUCTION:

Automation of tasks that are generally considered to require intelligent behavior for their accomplishment is a long-term goal of science and technology with relevance to numerous areas. The accomplishment of such tasks requires the design and construction of devices that can adapt to a possibly changing environment, in the sense of being able to produce behavior that can be modified in light of prior experiences. Modification of the device's behavior should be a consequence only of the interaction between the device and its environment.

It is only recently that enough fundamental scientific work, particularly in cybernetics, cognitive psychology and neurophysiology, has accumulated [1,2,3,4] that we can begin to understand the difficulties to be overcome in realizing this possibility.

Our present work involves the construction, in simulation, of networks of adaptive components capable of perceiving, remembering, and responding to stimuli from a simulated environment. Our interest is in embedded learning situations, i.e. ones in which the network's environment is an important part of the simulation. Fundamental questions include: How does one design an environment in which a given network is likely to function? How do we assess an environment in order to discover the parameters that will drive the design of a network which is likely to exhibit useful behavior in it? Can one find useful ways to model the component's adaptive mechanism? If there are several viable adaptive mechanisms, how does one choose the optimum mechanism to incorporate into the component's design? What topological features of the network are necessary and/or sufficient to predict its behavior when it is placed in a suitable environment?

The embedded network approach to be discussed here complements autonomous network simulations currently in progress at the Avionics Laboratory under the direction of A. Harry Klopf. Klopf's simulations are intended to provide a general model for the study of component-component and network-environment interactions. The intent of this ten-week study was to investigate particular special cases that would illuminate the relationship of Klopf's work to results that have recently been reported by Sinclair.[2]

II. OBJECTIVES:

The objective of this project was to establish a preliminary estimate of the relative capabilities of several adaptive net models in the accomplishment of closed-loop learning tasks. In order to achieve closure in the ten-week project period, it was decided to design a small simulation "test bed" that would function in the following way. This software, written in FORTRAN IV, consists of a simple environment and a small adaptive network. By employing various component modification algorithms, the behavior of the network in its environment can be observed, and we can arrive at a comparative assessment of the behavioral properties of networks vis-a-vis the modification algorithms.

The approach adopted toward the attainment of these goals, namely, the design of special cases of networks for comparative assessment, required the completion of a number of subtasks:

- Review Sinclair's book[2]
- Design neural modification algorithms for adaptive components including both all-excitatory and mixed excitatory-inhibitory versions of models based on

- the use principle
- the rest principle
- the heterostatic principle

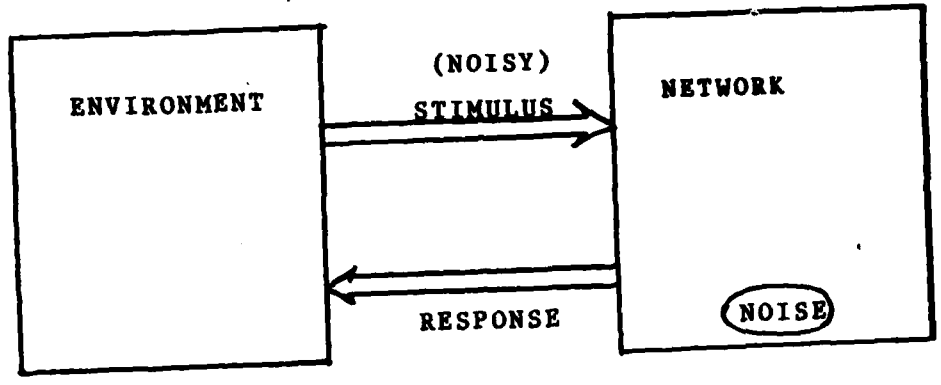
- Verify selected assertions of Sinclair and assess relative capabilities of the models
- Design, test and evaluate increasingly more complex networks and environments.

III. OUTLINE OF SIMULATION:

Figure 1 shows schematically the relation between an environment and an adaptive network contained in it. The situation depicted will be familiar to students of control theory, with one important change in point of view: we are interested here in adaptive control situations and not in the (more usual) steady-state, or nonadaptive, case.

At each step in (discrete) time, the environment presents to the network a stimulus which is a function solely of the state of the environment. In particular, the environment has no "memory" of previous states, nor does its state space change in time. The network receives the stimulus and computes a response, which alters the state of the environment, and this in turn causes the presentation of another stimulus to the network. The network contains a "memory" which is distributed over and indeed consists only of the strength of the connections between its adaptive components, and these connections are modifiable. Learning is equated with changes in the strength of these connections. The network's response is a function of the present values of connection strengths and a term representing internal noise in the components.

The environment, which is a one-dimensional version of the landmark-and-target environment of Sutton and Barto,[4,5] consists of a sequence of discrete positions numbered 0-10 from left to right that can be occupied one



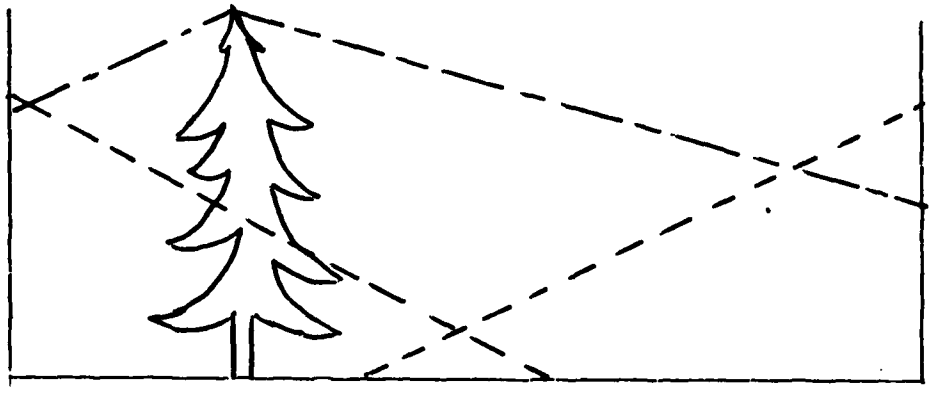
ENVIRONMENT PRODUCES A STIMULUS AS A FUNCTION OF ITS PRESENT STATE. NETWORK PRODUCES A RESPONSE WHICH ALTERS THE STATE OF THE NETWORK. IF SUBSEQUENT STIMULI ARE OF INCREASING "PREFERENCE," THE NETWORK ADAPTS TO INCREASE THE PROBABILITY THAT THE SAME RESPONSE WILL BE PRODUCED FOR THE SAME STIMULUS. NETWORK MUST HAVE SHORT-TERM MEMORY OF A TIME OR SPACE DIFFERENTIAL

Figure 1 - Adaptive Control Schematic

at a time by the network. At each end there is a "wall" beyond which the network cannot move. The environment also contains a fixed target, shown as a tree in Figure 2. The target and the walls are sources of signals that constitute the network's stimulus vector. Signal strengths decrease linearly with distance from the source and have maximum values that do not change during the simulation. The signals appear distinct to the network: it is not inappropriate to picture them as distinct odors or as light waves of separate wavelengths. Each signal has its own receptor in the network.

A typical network, shown in Figure 3, consists of a number of adaptive components that function in a way analogous to individual neurons or groups of neurons in an animal brain. It is convenient to employ the language of physiology here, bearing in mind that we do not intend our components to model any particular neurons that may be found in nature.

In particular, a component consists of a cell body (shown as a circle in Figure 3) and an axon, or communications channel, by means of which information is transmitted unidirectionally from the cell body to other cell bodies. Information passes from the axon to subsequent cell bodies across synapses, which are of various transmittances, or weights. Associated with each component is a firing threshold. The cell body fires when the membrane potential, which is the sum of incoming signals weighted by synaptic weights, exceeds threshold. The cell body, at any time, is in one of three states: firing (corresponding to neuronal depolarization sufficient to cause an action potential and subsequent axonal spike propagation in neurons); resting; or hyperpolarized. (Klopf's asymmetric heterostat discussed below requires the distinction between rest and hyperpolarization.)



LEFT
BOUNDARY

RIGHT
BOUNDARY

BOUNDARY STIMULUS NON ZERO FOR ABOUT HALF THE ENVIRONMENT, DECREASING LINEARLY WITH DISTANCE FROM BOUNDARY. TARGET STIMULUS STRENGTH ALSO DECREASES LINEARLY WITH DISTANCE FROM TARGET.

Figure 2 - A Simple Environment

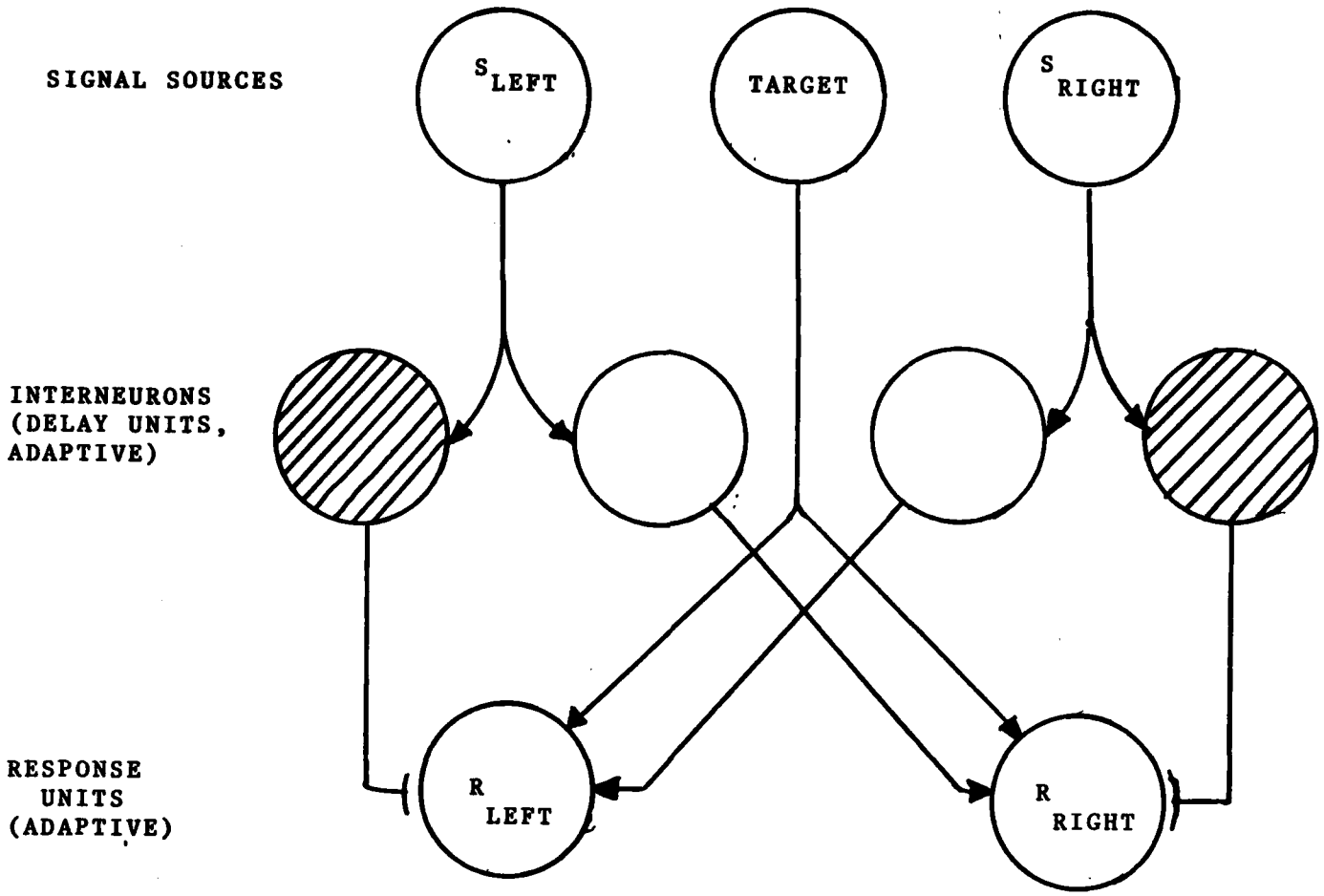


Figure 3 - A Target-Seeking Network With Inhibition - No Specialized Inputs

Components may be either inhibitory (shaded circles or excitatory. An inhibitory synapse, one driven by an inhibitory neuron, has a negative weight in the range -1.0 to 0.0 and excitatory synapses have weights between 0.0 and 1.0. Signal propagation time across one component is assumed to take one unit of time. At each time step the graded output of each component is computed and the network responds by firing either its left or right response component. If both membrane potentials exceed threshold, the component with the smaller of the two potentials is inhibited, so a definite move left or right results, unless the potentials are equal, in which case no movement is produced. Hence at each step the network will either move one step left or right or remain fixed.

Also at each step in time the network is receiving present (undelayed) target stimulus and past (delayed) wall stimuli. Since all stimuli are analogues of position, we consider the network to be receiving time-differential-of-position information. Since each movement of the organism results in increased or decreased target stimulation, the network can use this differential to modify its synaptic strengths, increasing the probability that it will move in the "right" direction (i.e. toward the target) the next time it is in the same position.

The simulation proceeds for a fixed number (currently 480) time steps and produces a graph of network position against time as well as some statistics regarding position and distance from target, averaged over all 480 steps.

IV. SYNAPTIC MODIFICATION:

In 1949, D. O. Hebb[3] proposed a hypothesis to explain how a synapse could be modified. His hypothesis has been nearly universally accepted by brain modelers, uncritically, we feel. Hebb's hypothesis is described by Sinclair[2] as the "use principle": synapses become stronger the more they are used. According to the use principle, an excitatory synapse becomes more efficacious as there is more correlation between pre- and postsynaptic activity. Hebb's use principle did not initially apply to inhibitory neurons, for they were not known at the time. Early simulations showed, however, that use principle networks were incapable of functioning unless large amounts of inhibition were present.

An objection to the use principle is that it is merely correlational, and does not take consequences into account. That actions have consequences in the form of feedback is a fundamentally important cybernetic insight, which ought to receive expression in any synaptic modification hypothesis that hopes to have broad explanatory capabilities. In fact, two such hypotheses (or rather principles that generate families of hypotheses) have recently been enunciated. Sinclair's[2] rest principle roughly says that synaptic efficacy increases if a synapse is first active in firing a cell and the cell is then allowed to rest, while repeated use weakens efficacy. The rest principle is thus a polar opposite of the use principle. The classical biological principle of homeostasis likewise has a polar opposite which was given expression by Klopf[1] as a synaptic plasticity hypothesis called heterostasis. The heterostatic synapse's efficacy changes in a way that reflects whether the cell as a whole receives excitation or inhibition subsequent to activity, so that activity leading to subsequent excitation becomes more likely to be repeated, while activity leading to subsequent inhibition becomes less likely to be repeated. Both

the rest and heterostatic principles seem to possess great explanatory power, and it is therefore of interest to compare them. For both principles, however, there is a spectrum of possible models, of which only two were studied in this project. First it is necessary to describe a feature of these models that sets them apart from use principle models.

ELIGIBILITY

According to either principle, not every synapse can be modified at any time: a synapse must first become eligible for modification through meeting some prior conditions involving pre- and postsynaptic activity. In this way, time, and possibly causality, are built into synaptic change. Sutton and Barto[4] review some of the reasons why this is a necessary feature in any hypothesis of learning that is to have adequate explanatory power.

Eligibility conditions in our models take the form of a 2-by-3 Boolean array, shown in general in Figure 4 and as a specific example in Figure 5. To be eligible requires that a synapse must exhibit, depending on the model, some of: (i) presynaptic and postsynaptic activity; (ii) presynaptic but not postsynaptic activity; (iii) post- but not preactivity; (iv) neither pre- nor postactivity; (v, vi) preactivity or inactivity together with postsynaptic hyperpolarization. In Figure 4, the rows refer to presynaptic activity, postsynaptic activity and hyperpolarization while the columns refer to our interest in the condition and then specifically its presence or absence. In Figure 5 are shown as examples the criteria for eligibility and modifiability of an excitatory synapse in a model Klopf calls the asymmetric heterostat. In this model (which we describe fully in following paragraphs), an

	Interest	Presence/Absence
Presynaptic activity		
Postsynaptic activity		
Postsynaptic hyperpolarization		

Figure 4 - Scheme for Defining Synaptic Eligibility and Modifiability. See Text for Explanation

	Condition Interested	Condition Present
Pre-activity	T	T
Post-activity	T	F
Post-hyperpolarization	F	-

Eligibility

	Interested	Condition Present
Pre-activity	F	-
Post-activity	T	T
Post-hyperpolarization	F	-

Modifiability

Figure 5 - An Example of Eligibility and Modifiability: Excitatory Side of an Asymmetric Heterostat

excitatory synapse becomes eligible for modification if its presynaptic side is active and its postsynaptic side is not; it is subsequently modified (increased efficacy) if subsequently but not too much later the postsynaptic cell becomes active. The increase in efficacy will act to move up the time at which the postsynaptic cell fires. In this model, we are not interested in hyperpolarization of the postsynaptic cell so row 3, column 2 entry is a "don't care" condition.

The reader may note that these schemes contain an asymmetry - there is no provision for presynaptic hyperpolarization. This is natural, for it is not a condition that can be detected at the synapse and does not therefore enter into the local synaptic situation.

TWO MODELS

The complete specification of one neuron will require four schemes: eligibility and modifiability must be specified for both excitatory and for inhibitory synapses. These in general will be distinct. Figure 6 displays the complete specification for two models: a rest principle model (first row of tables in Figure 6) and, below it, the asymmetric heterostat. The rules that specify these models follow.

REST MODEL

An excitatory synapse becomes eligible if both presynaptic and postsynaptic cells are active. An eligible excitatory synapse is modified if the presynaptic cell is active and the postsynaptic cell rests (i.e. is inactive); we do not distinguish between mere inactivity and hyperpolarization.

EXCITATORY CRITERIA		INHIBITORY CRITERIA	
ELIGIBILITY	MODIFIABILITY	ELIGIBILITY	MODIFIABILITY

(REST PRINCIPLE)

T	T
T	T
F	-

T	T
T	F
F	-

T	T
T	F
F	-

T	F
T	T
F	-

(ASYMMETRIC HETEROSTAT)

T	T
T	F
F	-

F	-
T	T
*F	-

T	T
T	T
F	-

F	-
F	-
T	T

Figure 6 - Complete Specification of Criteria for Eligibility and Modifiability for Two Synaptic Plasticity Models

On the inhibitory side, a synapse becomes eligible if the presynaptic cell is active and the postsynaptic cell is not. An eligible synapse is modifiable when the presynaptic cell becomes inactive at the same time the postsynaptic cell is active. In all rest models, repeated use decreases efficacy, so if a cell is modified immediately after eligibility, the synapse becomes weaker. Strengthening occurs if a period of rest elapses between eligibility and modifiability. The amount of synaptic change is an inverted U shaped function of time elapsed since eligibility.

ASYMMETRIC HETEROSTAT

This model, developed recently by Klopf[7], is a refinement of the models discussed in his book[1]. This project is the first known to simulate the asymmetric heterostat.

On the excitatory side, eligibility requires presynaptic activity coupled with postsynaptic inactivity - the presynaptic cell is "trying" to drive the postsynaptic cell. If, subsequently, the postsynaptic cell fires, the eligible excitatory synapses are strengthened.

On the inhibitory side, similar eligibility criteria are used as for excitation - an inhibitory presynaptic cell "tries" to hyperpolarize an active postsynaptic cell. If subsequently the postsynaptic cell is hyperpolarized, the synaptic efficacy is increased.

As was true of the rest principle, the amount of synaptic change is an inverted U shaped function of elapsed time since eligibility; there is however no initial weakening supposed for heterostats as there is for rest models. Sample weight change functions are shown in Figure 7.

V. SIMULATION DATA

This section contains representative data drawn from a number of simulation runs.

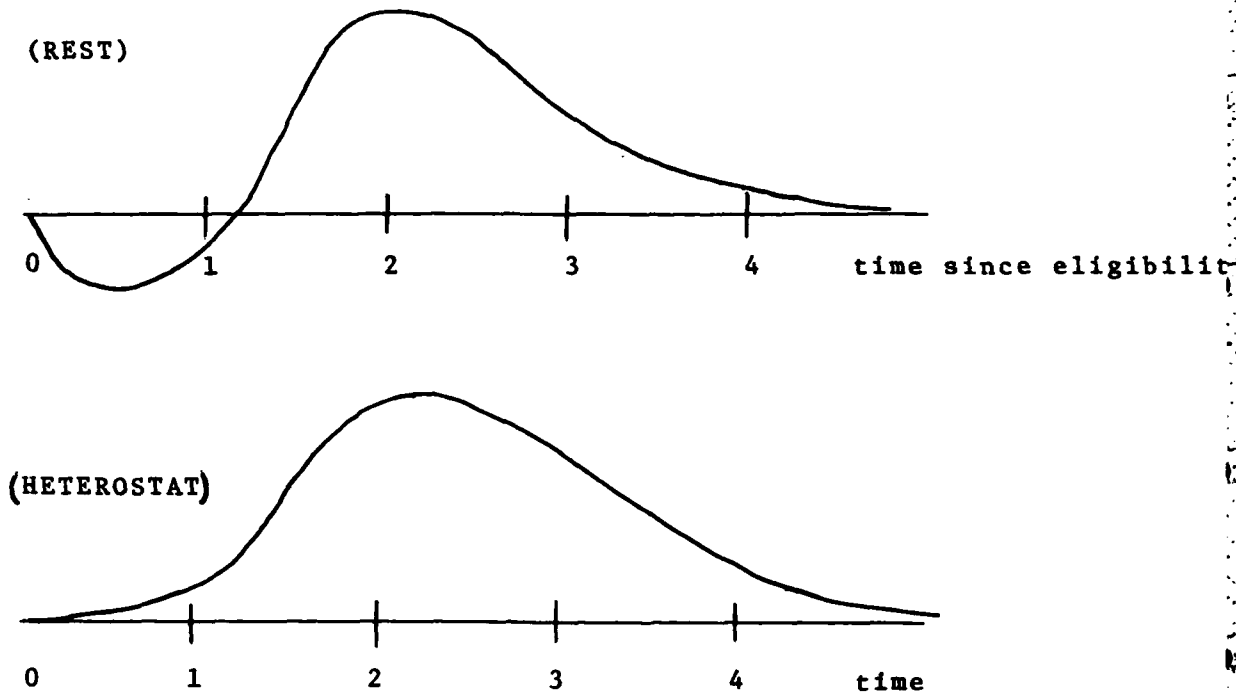


Figure 7 - Typical Weight Change Functions for Two Models

Figure 8 shows the simulation results obtained from several runs in which target position and initial network position were varied. The data are average network-to-target distance over 480 time steps. The array to the left shows the network's behavior in the absence of any target stimulus. On the right in Figure 8 are the average distances when the target is emitting a stimulus of 20 arbitrary units maximum. Each cell in an array contains two values: the upper value is the result using the rest model and the lower, the asymmetric heterostat. For example, with target position TPOS=1 and initial network position POS=1, the average distance to target is 4.0250 units in the absence of any target stimulus, and 3.4875 with a maximum target stimulus of 20.

Figure 9 shows a comparison of the results obtained when the simulation is run without allowing adaptation with results obtained when adaptation was allowed. In most cases it is usually the case that adaptation results in improved performance, as indicated by decreased average distances. Each cell in the nonadaptive case contains only a single number because the networks used were identical except for the adaptation algorithms.

VI. DISCUSSION:

Several of the subtasks described in Section II above were carried out in parallel.

First, regarding Sinclair's book:[2] this stimulating but ultimately frustrating work contains reports of many computer simulations, but is very sketchy. Some simulations constructed by us essentially verified the claims in Chapter 2 of the book. Later assertions did not fare so well, and the feeling we are left with is that many of Sinclair's assertions are supported by argumentations so qualitative that attempts to do numerical simulations will only lead to equivocal results, either supporting or refuting

TMAX=0

TPOS	POS 1	3	5	7	9
1	4.0250	4.0375	4.0458	4.0500	4.0625
	3.9833	3.9750	4.0083	4.0000	4.0250
3	2.0500	2.0375	2.0458	2.0500	2.0625
	1.9875	1.9750	2.0083	2.0000	2.0250
5	0.8625	0.8583	0.8541	0.8583	0.8791
	0.8917	0.8666	0.8708	0.8666	0.8875
7	1.9541	1.9625	1.9541	1.9541	1.9625
	2.0166	2.0250	1.9916	2.0041	2.0000
9	3.9541	3.9625	3.9541	3.9500	3.9416
	4.0166	4.0250	3.9916	4.0000	3.9791

TMAX=20

TPOS	POS 1	3	5	7	9
1	3.4875	3.5333	4.3958	4.0583	3.6250
	4.3000	4.2458	4.9875	4.0708	4.2250
3	2.1250	2.4375	2.4333	1.7833	1.6000
	2.2958	2.2750	2.1250	2.0708	2.2375
5	0.7875	0.8166	0.8125	0.7791	0.8083
	0.9166	0.9125	0.9583	0.9541	0.9250
7	2.5166	1.6000	2.2125	2.1667	2.3541
	1.7166	1.7416	1.9125	1.9041	1.7791
9	3.9916	3.6375	3.6291	4.1541	4.3750
	3.7416	3.7541	3.8750	3.9166	3.7625

Figure 8 - Average Network-to-Target Distance, as a Function of Target Position TPOS and Initial Network Position POS, after 480 Time Steps

TARGET POSITION	INI. NET POSITION	NONADAPTIVE	ADAPTIVE
9	1	3.9541	3.9916 3.7416
7	3	1.9370	1.6000 1.7416
5	5	0.8541	0.8125 0.9583
3	7	2.0750	1.7833 2.0708
1	9	4.0750	3.6350 4.2250

Figure 9 - Nonadaptive vs Adaptive Behavior Summary: Average Network-to-Target Distance After 480 Time Steps. Max Target Stimulus = 20, Max Wall Stimulus = 6 .

his assertions in accordance with the values assigned to various parameters. It is not possible at this time to make an assessment of the rest principle, partly because at least several hundred fundamentally different simulation models can be constructed on its basis. Sinclair himself admits 320 different possibilities, though his tabulation seriously undercounts the number of possibilities while simultaneously containing what appear to be logical inconsistencies.

We have successfully designed a single package of subroutines which can implement a large number of rest principle and heterostatic modification algorithms. The modification routine MODIFY is tailored to a specific algorithm by passing to it the appropriate eligibility/ modifiability criteria. MODIFY is supported by a number of short eligibility functions which are tailored to the individual models being simulated. The design of the software is itself an important outcome of this project, for it will be used in future studies. The main simulation will likewise be useful, for it was designed as an easily modifiable "test bed" capable of containing a broad range of more or less complex environments and networks. Local facility constraints required that coding be done in FORTRAN, rendering modifiability/maintainability of the software much lower than would be the case had the code been written in a modular language like Pascal or Ada. Time limitations have permitted the design and testing of only a small number of environments and networks. For most of the simulations, a decision was made to utilize the environment described earlier in Section III (see Figure 2) and the network shown in Figure 3.

The data obtained show clearly that this network is capable of responding to the presence of a target in its environment and to seek it out, at least weakly. It is clear in the perspective of hindsight however, that there is too much inhibition in the network. The network seems unable to

approach the walls at all closely, and seems to spend most of its time oscillating in the center of its range. This was not predicted, partly because we have not accumulated sufficient experience in designing and matching a network and environment. This represents one of the pioneering efforts in the area.

With due consideration of these design problems, the following conclusions can be drawn:

- Designing networks and environments that are mutually suitable is likely to be a complicated matter, for which there is at present no set of guidelines.
- Our network is able to adapt to the presence of a target in its environment and to seek it out; however:
  - the observed differences between rest principle and heterostatic components are not large enough to permit an assessment of their relative capabilities;
  - the network is on a plateau, responding to variations in parameters in only a small way.
- More information is needed on how an environment and an adaptive network interact:
  - we cannot yet draw conclusions on the relationships between large classes of adaptive elements and their environments;
  - the effects of introducing inhibition into a network are only beginning to be appreciated.

VII. RECOMMENDATIONS:

The experience gained in the design of adaptive networks and environments during the course of this ten weeks has been invaluable, but too brief to allow us to achieve all

of our objectives. Before a true assessment of the capabilities of rest models and heterostats can be carried out, two fundamental areas need to be mapped.

The first is to conduct a systematic study of the interactions between an adaptive network and its environment. The work done in this ten-week period has been a bare beginning, using ad hoc network designs and choices of environment. There is a need for a general study to shed light on the relationships between entire classes of adaptive networks and the environments in which they can function.

A second major study centers on the relative role of inhibition and excitation. These two kinds of inputs are not merely different in algebraic sign, but may play fundamentally different roles. Precisely how they differ needs to be determined.

In a slightly different direction, a useful study would be a systematic, point-by-point verification of the work reported by Sinclair in support of the rest principle as a neural plasticity hypothesis of broad explanatory power.

REFERENCES

1. A. Harry Klopf, The Hedonistic Neuron, Washington, D.C., Hemisphere Publishing Corporation, 1982.

2. John Sinclair, The Rest Principle: A neurophysiological theory of behavior, Hillsdale, New Jersey, Erlbaum Associates, 1981.

3. D. O. Hebb, Organization of Behavior, New York, Wiley, 1949.

4. R. S. Sutton and A. Barto, "Goal-Seeking Components for Adaptive Intelligence: An Initial Assessment," University of Massachussetts Final Report, Air Force Wright Aeronautical Laboratories Technical Report 81-1070 (AD A101476), Wright-Patterson AFB, Ohio, 1981.

5. R. S. Sutton and A. Barto, "Landmark Learning: An Illustration of Associative Search," Biological Cybernetics 42: 1-8, 1981.

6. R. S. Sutton and A. Barto, "Toward a Modern Theory of Adaptive Networks: Expectation and Prediction," Psychological Review 88 (No. 2): 135-170, 1981.

7. A. H. Klopf, personal communication, May, 1982.

1982 USAF-SCEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

AIRCRAFT ELECTRIC POWER GENERATION USING CASCADED SYMMETRICALLY WOUND MACHINES

Prepared by:	Dr. Thomas Ortmeyer
Academic Rank:	Assistant Professor
Department and University:	Department of Electrical and Computer Engineering Clarkson College
Research Location:	Air Forces Wright Aeronautical Laboratories Aerospace Power Division Power Systems Branch
USAF Research Colleague:	Dr. William U. Borger

Date:	July 23, 1982
Contract No:	F49620-82-C-0035

AIRCRAFT ELECTRIC POWER GENERATION USING CASCADED SYMMETRICALLY WOUND MACHINES

by

Thomas H. Ortmeyer

Brushless variable speed constant frequency electric power generation may be obtained using cascaded symmetrically wound machines. The feasibility of using these machines as the basis for a stand-alone aircraft generator system is investigated. The concept is attractive as the system operates without hydraulics and employs a solid state power converter which operates at a fraction of the system output power and frequency. Additionally, the converter has some degree of isolation from the load. These factors combine to offer a system of relatively low complexity with the potential for high reliability operation.

The operating characteristics of the two machines are examined. It is shown that, with proper choice of poles of the machines and with power factor correction of the generator load, the power handling requirements of the system can be minimized, at reasonable levels. Voltage and frequency control of the machines appears to be attainable without resorting to sensing internal machine variables. These results support a conclusion that the cascaded machines form the basis of a viable generator system. Recommendations for additional research on the system are included in this report.

Acknowledgement

The author would like to thank the Air Force Systems Command, the Air Force Office ot Scientific Research and the Southeastern Center for Electrical Engineering Education for the opportunity to participate in the Summer Faculty Research Program.

He would like to acknowledge the Air Force Wright Aeropropulsion Laboratory; and particularly the people of the Power Systems Branch for their hospitality during the course of the research period. He would also like to thank Dr. William U. Borger for suggesting this area of research, and for providing valuable insight and guidance throughout the research period.

I. Introduction

Aircraft electrical power is by necessity generated from a variable speed shaft. Additionally, these generators require brushless operation, and overall generator system weight is an important parameter in the design. There exist a variety of methods to provide constant frequency power from a variable speed source. (1) Systems presently in use, however, have lower reliability than is desired.

Aircraft generator systems can be divided into two types: those which mechanically convert variable shaft speed to constant shaft speed, and use a standard synchronous generator for mechanical to electrical power conversion; and, secondly, those which convert power from a variable shaft speed directly to electrical form. Recent developments in the areas of solid state power components and permanent magnet materials have led to new possibilities with the latter type of system; for example, a variable frequency generator coupled with a cycloconvertor to obtain constant frequency operation. (2) A machine with a set of symmetrical windings on both stator and rotor presents a method of obtaining constant frequency directly from a variable speed shaft. While this type of machine is generally used as a wound rotor induction motor, constant frequency generator operation may be obtained by strictly controlling the applied rotor electrical frequency. A variable voltage variable frequency power supply is required which delivers power to the rotor windings without brushes. While a variety of possibilities exist, the most promising option involves the use of a second symmetrically wound machine, as shown in Figure 1. This option minimizes the power handling requirements of the solid state converter with minimum system complexity.

The general nature and the operating characteristics of cascade connected symmetrically wound machines is investigated. It is concluded that such systems are feasible and would form a reliable nucleus for a variable speed, constant frequency generator system.

II. Objectives

The objective of this project is to determine the feasibility of using symmetrically wound machines as the basis of a variable speed constant frequency generator. Included in the project are three major steps:

(1) Determine system designs with the greatest potential for meeting the design requirements and goals.

(2) Establish specific ratings for system components, and for the excitation system.

(3) Identify voltage and frequency control methods and determine controllability of the system.

The scope of the present effort consists of analysis and simulations based on established machinery models which employ the "ideal machine" assumptions.

III. Steady State Operation

A symmetrically wound machine has frequency requirements

$$W_s = W_m \pm W_r \tag{1}$$

where all quantities are expressed in electrical rad/sec. W_s and W_r are the stator and rotor electrical frequencies and W_m is the mechanical speed. The addition or subtraction of rotor frequency from mechanical speed is governed by the phase sequence of the rotor currents.

With proper control of the rotor frequency, a constant frequency generator is obtained for a variable rotor speed. Neglecting the machine losses, the power flow in the machine is (3)

$$P_r = sP_s \tag{2}$$

$$P_m = (1-s)\, P_s \tag{3}$$

where s is the slip of the machine. The direction of these quantities is defined in the one line diagram of Figure 2.

In cases where it is desirable to minimize the power requirements of the rotor for a given speed range, the rotor must be driven through synchronous speed so that the power flowing into the rotor at minimum speed equals the power leaving the rotor at maximum speed. Under these conditions, the maximum power requirements of the rotor circuit are

$$\left(\frac{P_r}{P_s}\right)_{max} = \frac{Y-1}{Y+1} \tag{4}$$

where Y is the ratio of maximum speed to minimum speed.

IV. Cascaded Machines

To meet the brushless requirements for an aircraft system, two machines can be cascaded, as shown in the one line diagram of Figure 3. As direct current cannot be induced in the rotor windings, the power machine (m1) must be operated strictly in the subsynchronous mode ($s_1 < 0$) or the supersynchronous mode ($s_1 > 0$).

The power requirements of the field windings of machine 2 are

$$P_f = s_2 P_A = s_1 s_2 P_s \tag{5}$$

where;

$$s_1 = \frac{W_s - W_{m1}}{W_s} \tag{6}$$

$$s_2 = \frac{W_r - W_{m2}}{W_r} \tag{7}$$

Note that s_2 is the machine 2 slip referenced to the rotor. W_{m1} and W_{m2} are the rotor speed in electrical radians per second for machine 1 and machine 2, respectively, so

$$W_{m2} = (P_2/P_1)W_{m1} \tag{8}$$

where P_1 and P_2 are the number of poles for machines 1 and 2 respectively.

As the frequencies in the rotor circuits are equal,

$$W_A = W_r = s_1 W_s \tag{9}$$

Combining Eq's 5 through 9,

$$\frac{P_f}{P_s} = s_1 s_2 = s_1 \pm \frac{P_2}{P_1}(1-s_1) \tag{10}$$

The sign of the final term of Eq. 10 is governed by the phase connections between the r windings of machine 1 and the a windings of machine 2. These two connections will be referred to as the plus and minus rotor connections.

To minimize the excitation power requirements of the two machine system, machine 2 is driven through synchronous speed so that the power input to the field at minimum speed matches the power output from the field at maximum speed. Under these conditions, the maximum field power required is

$$\left(\frac{P_f}{P_0}\right)_{max} = \frac{Y-1}{Y+1} \tag{11}$$

Again, Y is the ratio of maximum speed to minimum speed, and it can be seen that the field power requirement for this system is the same as for the single machine system.

This power requirement can be met for any combination of poles on machines 1 and 2 with machine 1 operation either above or below synchronous speed and either the plus or minus rotor connection. Many of these combinations require excessive power levels on the rotor circuit, however.

The most promising configurations for a 1.5:1 speed range, 400 hertz generator are given in Table I. Note that the plus connection with $P_1 = P_2$ is not a useful connection, as has been previously noted. (4).

To evaluate the var effects and to include the effects of winding resistance, calculations were performed using typical machine values listed in Table II. The load power was considered to be inductive with power factor greater than 75 percent, and with machine 1 rated volt-amps equal to maximum load volt-amps. The volt amp rating of machine 2 is reduced by the ratio (Pr/Ps)max of Table I, and the frequency rating of this machine is set at maximum armature frequency. The volt-amp and watt requirements of the field winding are shown in Figure 4 for the five cases under study. Figure 4 also includes data with power factor improvement capacitors at the generator terminals. These capacitors are considered to be connected to the generator for all load conditions.

The results show that operating configuration as well as var control have an effect on field volt-amp requirements. The subsynchronous cases in general have larger volt-amp requirements than the supersynchronous cases. Only with full compensation do the field var requirements approach desirable levels.

The peak power and volt-amp flows in the rotor windings are shown in Figure 5. Again, the five cases are shown with no compensation, half compensation, and full compensation. The figure shows wide variation between cases, and also a marked influence of compensation on the rotor requirements.

Comparisons of this data shows no clearly superior case. The final decision would involve a trade-off between machine speed and weight, excitation supply requirements and overall system cost. The exciter must be a variable frequency, variable voltage source with voltage essentially proportional to frequency. It must be capable of both receiving and delivering power and must be able to smoothly go through zero hertz from forward to reverse rotation. The maximum frequency for a 400 hertz system with 1.5:1 speed range is 80 hertz for all cases. This value is well within present technology for both cycloconverter and PWM inverter sources.

Table I

Maximum power flow requirements for various system configurations. All machine losses neglected.

Case	Speed Range (rpm)	Poles P1	Poles P2	Operating Mode	Rotor Connection	$\left(\frac{Pr}{Ps}\right)_{max}$	$\left(\frac{Pf}{Ps}\right)_{max}$
A	9600-14400	2	2	sub-synchronous	minus	.60	.20
B	6400-9600	4	2	sub-synchronous	minus	.47	.20
C	4800-7200	6	2	sub-synchronous	minus	.40	.20
D	9600-14400	6	2	super-synchronous	plus	.80	.20
E	6400-9600	8	2	super-synchronous	plus	.60	.20

Table II

Typical motor parameters in per unit to the motor base.

Parameter		Per unit value
M_1	M_2	
r_s	r_a	0.02
r_r	r_f	0.025
x_s	x_a	0.08
x_r	x_f	0.08
X_m	x_{am}	4.0

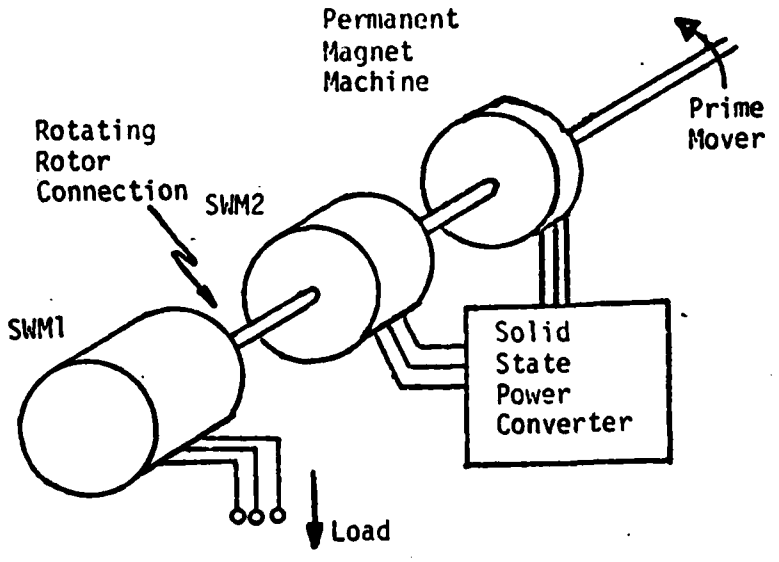


Figure 1. Stand-alone brushless variable speed constant frequency generator system using symmetrically wound machines (SWM).

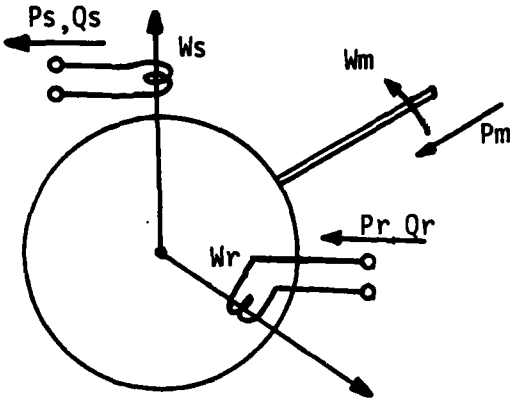


Figure 2. One line diagram of symmetrically wound machine showing generator convention power flow direction.

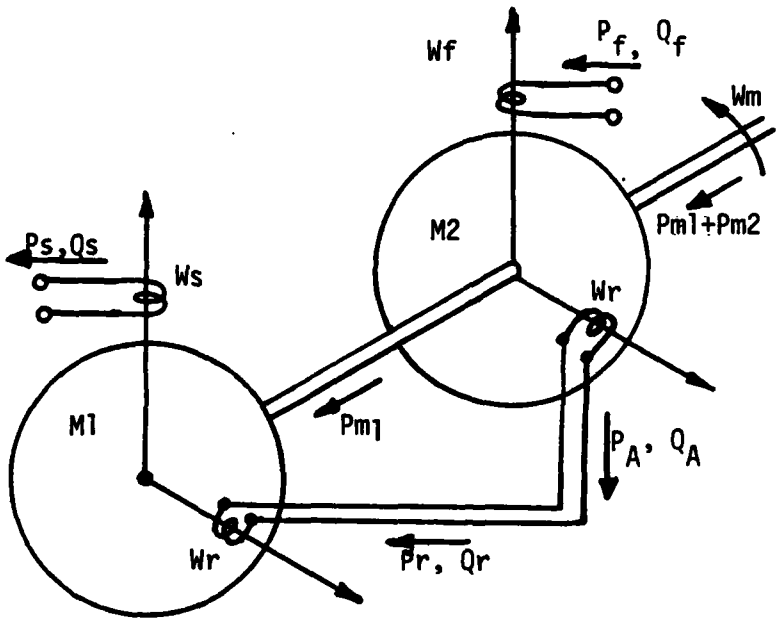


Figure 3. One line diagram of symmetrically wound machines showing generator convention power flow directions.

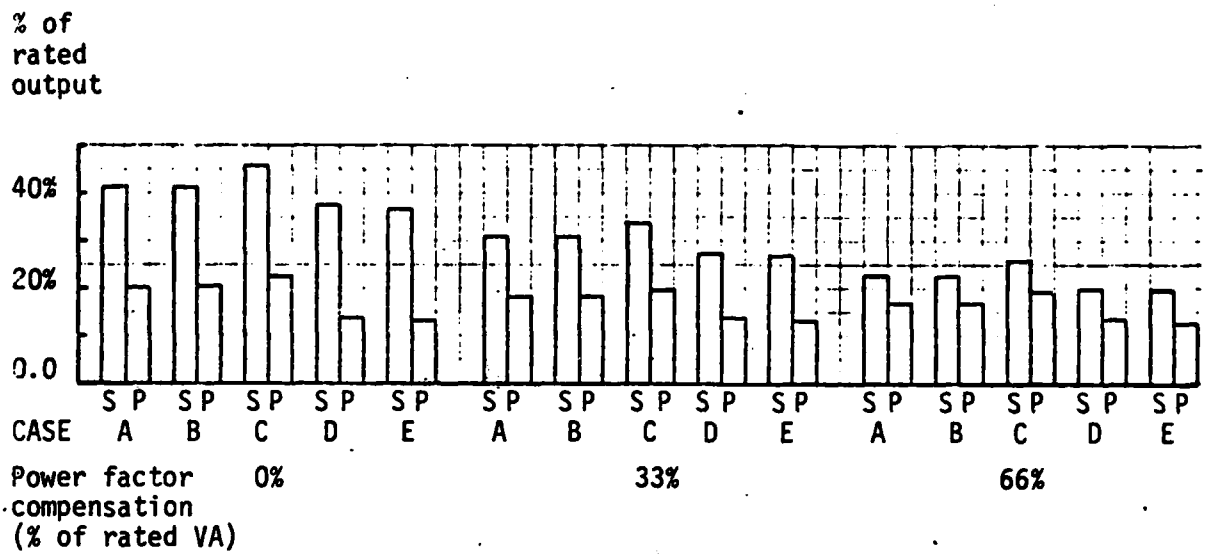


Figure 4. Maximum field volt amp (S) and power (P) requirements for 1.5:1 speed range generators.

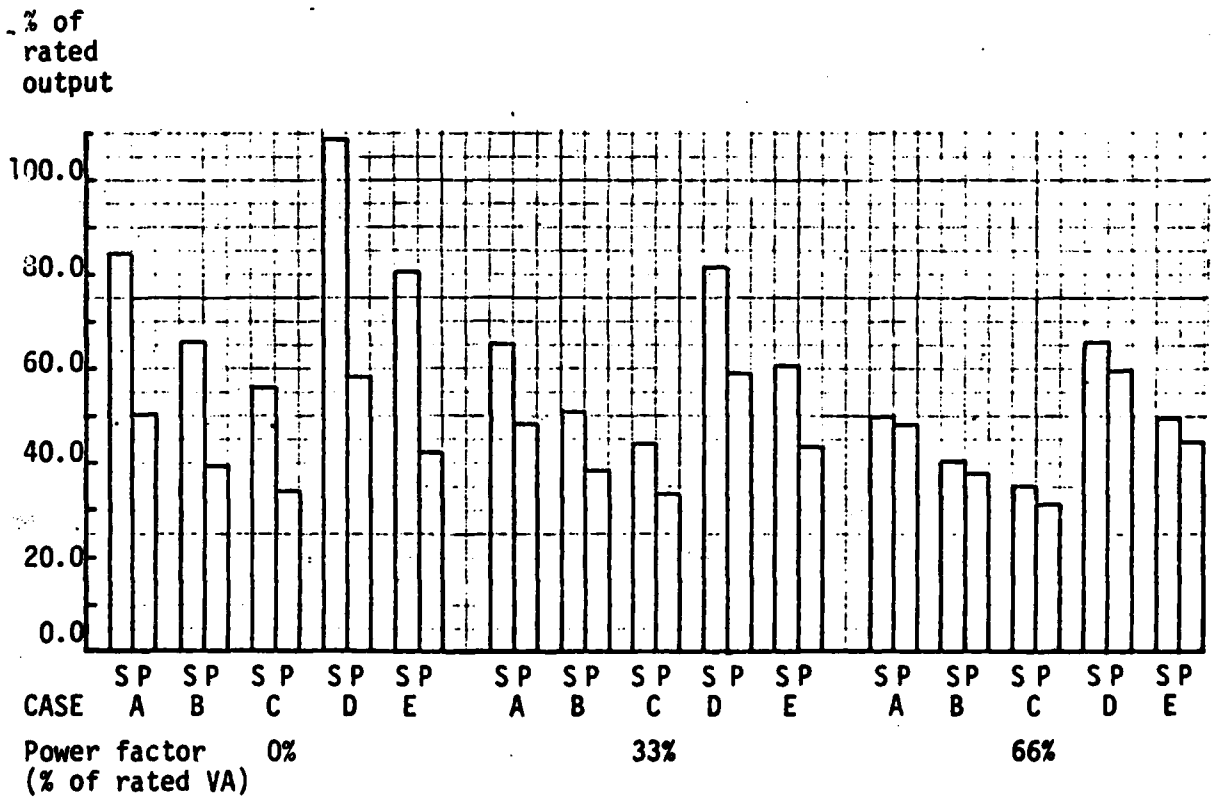


Figure 5. Maximum rotor circuit volt-amp (S) and power (P) requirements for 1.5:1 speed range generators.

V. Frequency and Voltage Control

Output voltage and frequency of the system must be maintained within specified limits. This control is achieved by adjusting the voltage and frequency applied to the field of the cascaded machines. The most direct control scheme is outlined in Figure 6. The field frequency is the difference between desired output frequency and the mechanical speed, properly scaled. A voltage error signal is fed to the controller, which in turn sends the field voltage command to the excitation source. For the purposes of the study, it is assumed that the excitation source is sinusoidal and that there are no unintentional delays in the control system. The shaft speed is considered to be variable but independent of the electrical system.

The most challenging aspect of the control of the system is the stabilization of the flux wave in the second machine, particularly as the excitation for this machine is swept through zero frequency with a reversal of the phase sequence. This particular subproblem can be analyzed independently by assuming that no output current flows from this machine. The linear incremental equations of the machine are then, in per unit and transformed to the frequency domain,

$$\begin{bmatrix} v_{df1} \\ v_{qf1} \end{bmatrix} = \begin{bmatrix} r_f + sL_f & -W_{s20}L_f \\ W_{s20}L_f & r_f + sL_f \end{bmatrix} \begin{bmatrix} i_{df1} \\ i_{qf1} \end{bmatrix} + \begin{bmatrix} -\lambda_{qf0} \\ \lambda_{df0} \end{bmatrix} W_{s21} \qquad (12)$$

$$\begin{bmatrix} v_{da1} \\ v_{qa1} \end{bmatrix} = \begin{bmatrix} sL_{AM} & -W_{e20}L_{AM} \\ W_{e20}L_{AM} & sL_{AM} \end{bmatrix} \begin{bmatrix} i_{df1} \\ i_{qf1} \end{bmatrix} + \begin{bmatrix} -\lambda_{qa0} \\ \lambda_{da0} \end{bmatrix} W_{e21}$$

a and f refer to the armature and field windings of the machine, W_{e2} is the armature frequency, and W_{s2} is the slip frequency of this machine.

Subscript 0 refers to steady state values and 1 refers to incremental values. (5)

In Equation 12, the machine is represented by d and q axis rotating in synchronism with the field voltages. V_{qf1} can be chosen to equal zero without loss of generality. The transfer function of most interest for the machine is output voltage magnitude. The incremental output voltage magnitude is

$$V_{a1} = \left(\frac{V_{da0}}{V_{a0}}\right) V_{da1} + \left(\frac{V_{qa0}}{V_{a0}}\right) V_{qa1} \qquad (13)$$

and the transfer function is then

$$\frac{V_{a1}}{V_{f1}} = \frac{L_{am}}{(r_f^2 + W_{s20}^2 L_f^2)^{1/2}} \quad \frac{s^2 W_{s20} + s\, W_{e20}(r_f/L_f) + W_{e20}(W_{s20}^2 + (r_f/L_f)^2)}{s^2 + 2s(r_f/L_f) + (r_f/L_f)^2 + W_{s20}^2} \qquad (14)$$

Items of interest in Eq. 14 are:

° The poles are located at $s = -r_f/L_f \pm j\, W_{s20}$

° The effective system gain varies widely with slip frequency, and changes sign as slip goes through zero.

° The system zeroes are in the left half plane for positive slip.

° One zero is located in the left half plane and one in the right half plane when slip is negative.

Feedback of the output voltage with an integral controller is shown in Figure 6, with $G_1(s) = G_1/s$. Routh's criterion shows that the system is

stable at any pair of frequencies W_{s20} and W_{e20} if

$$G_1 < \frac{2\left[(r_f/L_f)^2 + W_{s20}^2\right]^{\frac{1}{2}} r_f}{(W_{e20} - W_{s20})\, L_{AM}} \qquad (15)$$

A value for G_1 stable at all frequencies of interest may then be found. Studies have shown, however, that this value will produce sluggish response at high slip operation.

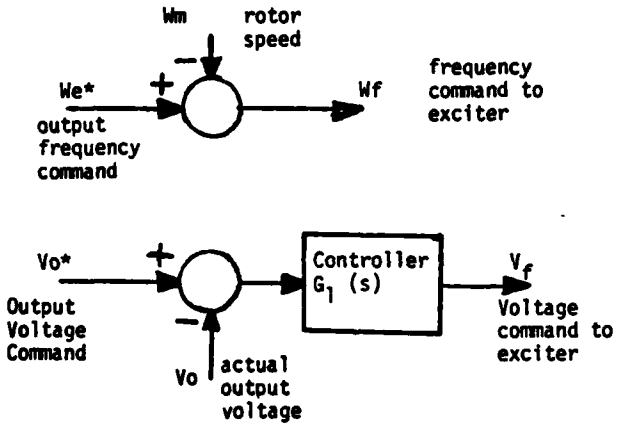


Figure 6. Basic control system using speed and voltage sensing

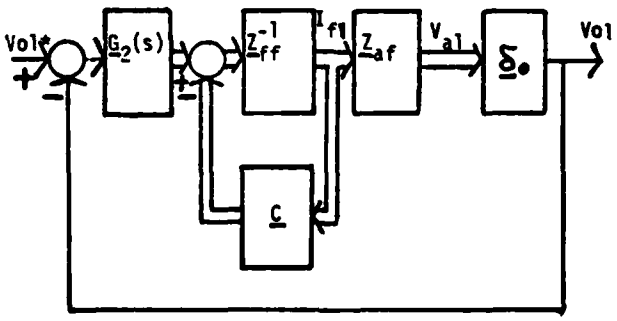


Figure 7. Incremental block diagram showing voltage controller $\underline{G}_2(s)$ and state feedback matrix $\underline{C}$.

VI. Possibilities of State Feedback

Figure 7 is a state variable representation of Equations 12 and 13, where $V_{f1} = \begin{bmatrix} v_{df1} & v_{qf1} \end{bmatrix}^T$ $I_{f1} = \begin{bmatrix} i_{df1} & i_{qf1} \end{bmatrix}^T$ and $V_{a1} = \begin{bmatrix} v_{da1} & v_{qa1} \end{bmatrix}^T$

The compensation matrices are $G_2 = \begin{bmatrix} K/s & 0 \end{bmatrix}^T$ and the state feedback matrix C which will be chosen to preserve the symmetry of the system,

$$\underline{C} = \begin{bmatrix} c_1 & -c_2 \\ c_2 & c_1 \end{bmatrix} \tag{16}$$

Reduction of the state feedback loop shows that the system poles with the outer loop open are located at

$$s = -\frac{r_f + c_1}{L_f} \pm j\left(\frac{c_2}{L_f} + W_{s20}\right) \tag{17}$$

The closed loop poles can, therefore, be placed nearly arbitrarily.

Simulations of the machine show that the linearization is valid and that the addition of load to the machine does not markedly affect system response. The simulations were conducted assuming sinusoidal conditions and a series R-L load.

VII. Stability of the Two-Machine System

A linearized analysis of two cascade connnected symmetrically wound machines was conducted with an integral controller regulating the output voltage of the system and with the possibility of state feedback compensation involving the field currents only. The particular system studied is case C of Table I, with machine parameters as listed in Table II.

Root loci for the system without state feedback are shown in Figure 8. The upper half of the s plane is shown, with the dominant poles included.

Each plot shows the effect of the integral controller gain on the dominant closed loop poles of the system at machine speeds of interest. Analysis of Figure 8 shows an open loop pole near the rotor frequency, with a zero located near this pole. A second pole exists near the frequency of the machine 2 field.

Both poles are lightly damped, and the pole associated with the field winding drifts into the right half plane with low values of gain.

By implementing state feedback with $c_1 = 59r_f$ and $c_2 = 0$ from Equation 16, the root loci of the system are shifted as shown in Figure 9. State feedback moves the field pole away from the imaginary axis, which significantly improves motor performance. The system remains stable over the entire speed range for a wide range of gains, with instability only under high speed, high gain conditions.

System loading has an important, although secondary, influence on the system performance. As expected, the load provides additional system damping, and the worst case from a stability viewpoint occurs under lightly loaded conditions.

These studies show that the cascaded machines are controllable with a reasonably straight-forward control strategy. In particular, the difficult to sense values of internal flux and rotor currents do not appear to be necessary for control. Additional studies have shown that generators operating in the supersynchronous mode is somewhat more involved, and more elaborate compensation may be required in ceratin cases of supersynchronous operation.

VIII. Recommendations

The use of symmetrically wound machines as variable speed, constant frequency generators has been investigated. These machines show good potential for use in a variety of applications.

The brushless operation required for aircraft systems is best achieved by cascading two machines. The cascading allows minimization of the excitation power requirements while also limiting the system complexity.

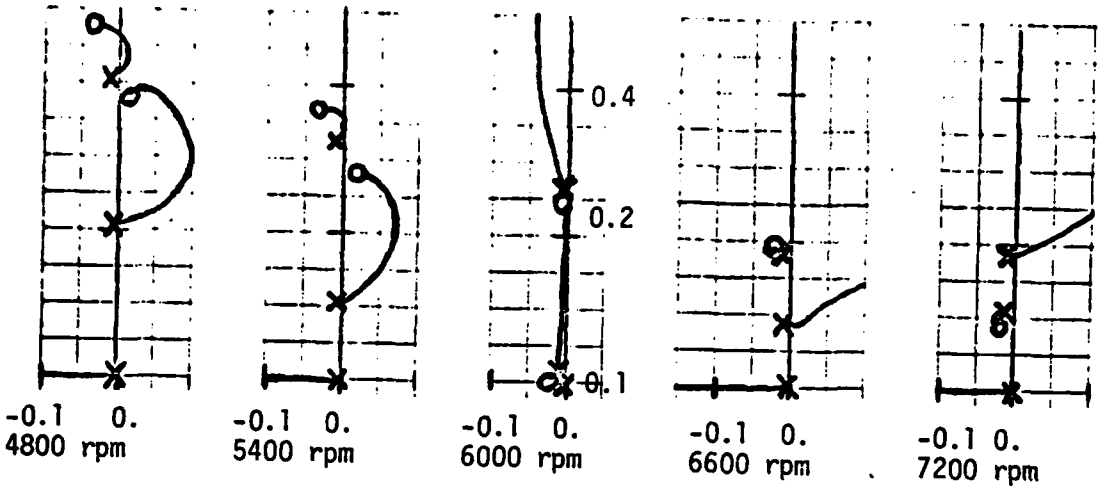


Figure 8. Root Locus of Vo/Vo* with integral controller at various shaft speeds. Values in per unit with $S_{BASE} = 2\pi 400$.

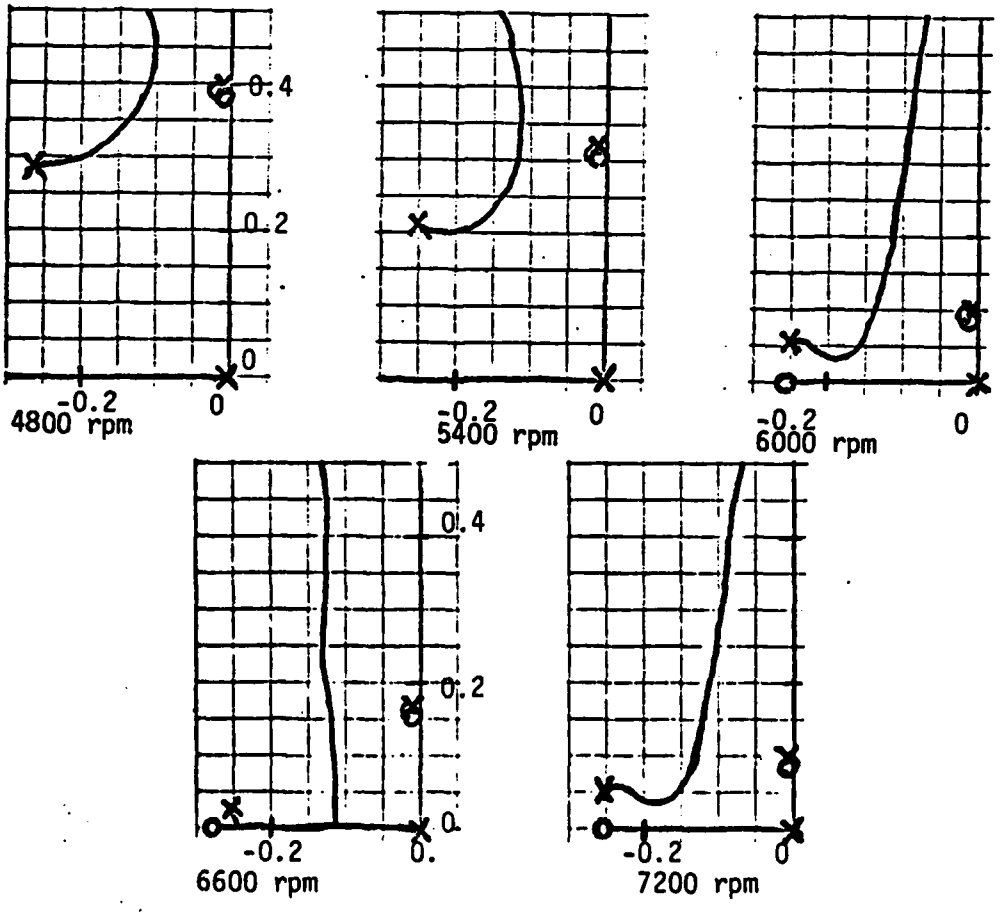


Figure 9. Root locus of Vo/Vo* with integral controller and field current feedback at various shaft speeds. Values in per unit with $S_{BASE} = 2\pi 400$.

The resulting two machine system forms a brushless variable speed constant frequency generator with excitation requirements which increase as the speed range to be covered increases.

Detailed analysis shows that excitation requirements and the power handling requirements on the rotor can be maintained within reasonable limits with proper choice of speed range and poles for the two machines, and with the use of power factor correction capacitors at the output of the generator.

The study shows that the resulting system is controllable and will provide good response. The control system is not overly complex and will not require sensing of rotor currents or internal machine fluxes.

The basic feasibility of the system has been shown. To bring the system to practical application, however, will require further investigations. Two basic areas exist which have been beyond the scope of the present investigation. The first involves the performance of the cascaded machines under nonideal conditions. The ability of the machines to meet military specifications with unbalanced loads and system faults is critically important. Additionally, the effect of nonsinusoidal excitation voltages on the generator output must be understood and controlled. Very little knowledge exists at present in these aspects of the machines performance.

The second area of interest involves the excitation system for the generator. While the requirements of the system have been defined by the present study, the optimum system configuration has not been determined.

Basic question which must be addressed include:

(1) The use of cycloconverter or rectifier-inverter

(2) Triggering strategies, particularly at and near the zero frequency point.

(3) The use of permanent magnet machine or wound rotor machine to provide excitation power.

These steps are intermediate to the goal of building a high reliability, low weight variable speed constant frequency generator system. The results of the present study indicate a good likelihood that a generator system using symmetrically wound machines would offer reliable, high quality electrical power, at competitive system weight and cost.

References

1. T.B. Owen. "Variable-Speed Constant-Frequency Devices: A Survey of the Methods in Use and Proposed." AIEE Trans. vol. 78 pt. II, 1959. pp. 25-30.

2. W. U. Borger. "A 400 Hz Starter/Generator System" Proceeding of the IEEE 1982 National Aerospace and Electronics Conference, May 18-20, 1982. pp. 1031 - 1036.

3. M. Riaz. "Energy-Conversion Properties of Induction Machines in Variable-Speed Constant-Frequency Generating Systems." AIEE Trans. Vol. 78 pt. II, 1959. pp 25 - 30.

4. E. Erdelyi and E. E. Kolatrowicz. "The Limitations of Induction Generators in Constant-Frequency Aircraft Systems." AIEE Trans. vol. 77 pt. II, 1958. pp. 348 - 52.

5. A. E. Fitzgerald and C. Kingsley, Jr. Electric Machinery, second edition, 1961. New York: McGraw-Hill.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

ADAPTATION TO OPTICAL FLOW RATES DURING LOW ALTITUDE, HIGH SPEED FLIGHT

Prepared by:	Dr. Dean H. Owen
Academic Rank:	Associate Professor
Department and University:	Department of Psychology The Ohio State University
Research Location:	Air Force Human Resources Laboratory, Operations Training Division, Williams Air Force Base
USAF Research Colleague:	Dr. Thomas M. Longridge
Date:	August 20, 1982
Contract No.	F49620-82-C-0035

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

A PRELIMINARY STUDY OF ELECTROMIGRATION IN ALUMINUM AND ALUMINUM-SILICON FILMS

Prepared by:	Dr. R. L. Remke Kevin Wilson
Academic Rank:	Assistant Professor Graduate Student
Department and University:	Dept. of Electrical and Electronic Systems, Univ. of South Florida
Research Location:	Rome Air Development Center, Rome, NY Reliability Branch, Product Evaluation Section
USAF Research Colleague:	Dr. Robert Thomas
Date:	September 10, 1982
Contract No:	F49620-82-C-0035

ADAPTATION TO OPTICAL FLOW RATES DURING LOW ALTITUDE, HIGH SPEED FLIGHT

by

Dean H. Owen

ABSTRACT

Research with a driving simulator demonstrated that drivers adapt to high rates of speed, increasing actual speed in order to hold apparent speed constant. Adaptation affected sensitivity to subsequent change in speed and the aftereffect persisted for up to 20 sec. If adaptation is to optical flow rate rather than speed per se, pilots can compensate for adaptation in two ways: (1) by increasing speed and/or (2) by losing altitude. Two preliminary experiments were conducted using the Advanced Simulator for Pilot Training to determine first whether pilots adapt to optical flow during low altitude, high speed flight. If it is shown pilots do adapt, further research will test for effects on sensitivity to changes in altitude and speed. Embedded within the adaptation period was a second experiment on sensitivity to lateral displacement. Time to initiate a rudder-pedal adjustment will be related to fractional increase in displacements simulating 90-deg crosswinds.

ACKNOWLEDGMENTS

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to spend a worthwhile and interesting summer at the Air Force Human Resources Laboratory, Williams AFB, AZ. He wishes to extend his appreciation to Steve Stephens, Chuck Snow, Brian Rogers, Jim Rosinke, and Tom Farnan, of the laboratory staff, and Don Buck, Tom Dickens, Joe Berfanger, and Lee Cole, of Singer-Link, for their cooperation, insights, and considerable efforts during the development of the research. He also acknowledges helpful discussions with and feedback from Drs. Elizabeth Martin and Joe DeMaio.

Finally, he would like to thank Dr. Thomas M. Longridge for his invitation to come to the laboratory, his ability to expedite the research effort, and his personal involvement in the project.

I. INTRODUCTION:

It is widely acknowledged that during low altitude, high speed flight the ground is a formidable enemy, and accident statistics support the claim. Even though aid is available from instruments, pilots find that the required tasks demand attention to information available outside the aircraft for reasonably long periods of time. Part of what a pilot must attend to is visual information specifying where he is relative to the ground, where he is headed and how fast, as well as what obstacles must be avoided, and when changes in flight path and speed must be initiated, modified, and terminated. Following the approach to the study of perception developed by Gibson[1,2], it is assumed that there is adequate information available in the optic array along a path of self motion to allow a pilot to safely control an aircraft.

In the Aviation Psychology Laboratory at Ohio State University, the author and his colleagues have isolated several sources of optical information useful in detecting changes in altitude and speed[3,4]. All of these sources are global transformations in the optic array, i.e., they specify rates of change common to all angular velocities in the array, hence, along all possible lines of regard[5].

As is true with most perceptual phenomena, however, apparent speed of self motion decreases over time even though environmental characteristics of an event such as altitude, speed, and direction of travel are held constant, i.e., adaptation occurs. Denton's[6,7,8] research on adaptation to self motion was conducted with a driving simulator in which eyeheight (altitude of the eye above the ground) was constant. In these situations path speed and global optical flow rate (path speed scaled in eyeheights) differ only by a scale factor[5]. When altitude is varied, the two can be separated, as in the familiar case during flight where one appears to be moving very slowly at high altitude but very rapidly close to the ground, even though ground speed is the same for both situations.

In the case of driving, the consequence of adaptation is one-dimensional: As the driver adapts, actual speed is increased to compensate for the reduction in apparent speed. But in the case of flying, the

pilot has two alternatives, since flow rate increases as eyeheight decreases: Either loss in altitude or increase in speed will compensate for the effect of adaptation. For flight, just as with driving, increased speed reduces the margin of safety as an obstacle is approached, but a compensating advantage in flight is increased lift at higher speed. Loss in altitude, however, simply reduces the margin of safety with no compensating advantages. If the optical acceleration which accompanies a constant descent rate goes unnoticed due to adaptation, and other optical changes (e.g., increase in perspectival splay and decrease in optical density) are subthreshold, the pilot could become dangerously low without being aware of the loss in altitude.

There are three compelling reasons for beginning with a test for an aftereffect of adaptation before moving on to the more interesting and critical issues of changes in sensitivity: (1) Even though Denton's experiments were psychophysically sound, his tasks all depend on subjective perceptual reports, i.e., he has no way to objectively score a response as correct or incorrect. Use of forced-choice tasks will remedy this problem. Furthermore, Denton did not have a control condition with which to compare the aftereffect of adaptation. Introduction of a pretest, for which no variation is predicted over adaptation flow rates, provides a standard against which to compare effects of adaptation conditions on the posttest. (In the first preliminary experiment with the flight simulator, it was concluded from posttest results alone that different flow rates during the adaptation period had produced different aftereffects. Since the same pattern of results occurred for the pretest where the flow rates could have no effect - all pretest reaction times should have been the same - the posttest result could not be taken at face value.

(2) There are many alternative explanations of why pilots fly too low and/or too fast, e.g., inattention, complacence, or just "pushing the limits." Adaptation may ultimately account for only part of the problem or perhaps even none. Therefore a series of experimental operations are required to assess the contribution of adaptation. The first is to show that pilots adapt to optical flow rates specifying self

motion. A standard criterion for adaptation is demonstration of a complementary aftereffect.

(3) Denton found and experimentally exploited extensive individual differences in adaptation to self motion. Although it is not clear from his papers what test he used, Denton selected from a larger sample 12 subjects who revealed essentially no motion aftereffect and 12 showing relatively large aftereffects. It is not known to what extent pilots in general and those flying at low altitude in particular are preselected on the basis of either susceptibility to adaptation or sensitivity to optical flow variables. Correlated susceptibility to motion sickness or insensitivity to information necessary to fly safely may result in self selection or elimination from training by instructors. Therefore it is essential to determine the extent to which Denton's findings with individuals selected from the extremes of a population of drivers generalize to a group of pilots whose characteristics relevant to the experimental tasks are completely unknown.

The primary goal of the research initiated during the fellowship period was to replicate an earlier adaptation study[8] within the context of flight simulation using experienced pilots as subjects. If adaptation to optical flow rates can be demonstrated during flight, then a program of research will be initiated to determine whether pilots' sensitivity to critical kinds of visual information changes following adaptation. A secondary goal was to explore the development of a research paradigm designed to study a pilot's perception within the context of interactive flight tasks.

In order to ensure that the pilots attended to the flow pattern during the adaptation experiment, they were given the task of staying on course above a white stripe extending toward the horizon ahead. In addition to keeping the pilot unaware that the experiment was primarily concerned with adaptation, the task provided the first opportunity to explore a set of interrelated theoretical assumptions.

Gibson[2,9] proposed that perception guides action and action makes available information for perception. From a different theoretical perspective, that of control theory, Powers[10] argued that the sole

purpose of behavior was to control perception. Working from these assumptions, Owen and Warren[11] have proposed that optical parameters be used as independent variables to initiate experimental events, that the individual's control actions be used as perceptual reports, and that the optical variables and invariants produced by the adjustments be measured in optical terms as dependent variables. This procedure allowed a test of sensitivity to experimenter controlled transformations in the optical flow pattern and provided an opportunity to determine the desired optical conditions the pilot is attempting to achieve by making control adjustments in response to cross wind displacements.

II. RESEARCH:

Two preliminary experiments have been conducted to check out the testing and data recording systems and to adjust experimental parameters. All aspects of the research are presented in detail, since they are anticipated to be the same in the final experiment. The design and procedure are divided into sections on the test for an aftereffect of adaptation and a test for sensitivity to lateral displacement embedded in the adaptation period.

Simulation facility. The Advanced Simulator for Pilot Training (ASPT) in the Air Force Human Resources Laboratory, Williams AFB, Arizona, was used for data collection. The ASPT facility has two major advantages relevant to the study of self-motion perception: (1) The seven-screen domed visual display system fills most of the visual field in the forward and lateral directions, so that many pilots actually experience self motion. The phenomenon is seldom experienced with small-field frontal displays. (2) The system is fully interactive, so that a pilot's control adjustments change the visual display appropriately to mimic visual information for self motion. The ASPT system, including visual displays and computer image generation, has been described in detail elsewhere[12]. Therefore, only aspects directly relevant to the current study will be described.

The simulator configured as an F-16 aircraft was used for testing. The motion base was not used, so that the only information available about self motion was visual. All instruments were covered.

The field of view is approximately ±150 deg horizontal by +110 deg and -40 deg vertical. The CRTs use PT-462 phosphor and appear monochrome green. Each CRT has a total radiated output in excess of 10,000 lumens, providing a display brightness at the pilot's view point of 6 ft lamberts. Each display has 1,023 scan lines, 1,000 elements/line, 7 arc-min resolution, and cycles at 30 frames/sec. Under interactive conditions, the image generation system makes transformations appropriate to the pilots control adjustment in less than 100 msec.

The environmental surface simulated consisted of horizontal bars alternating in light and dark "gray" levels parallel to the horizon directly ahead. The bars were 20,000 ft wide (in the y or lateral dimension) and 200 ft deep (in the x or front/back dimension), with the entire textured surface extending 34 mi in the x dimension. A "white" stripe 10 ft wide, perpendicular to the horizontal bars, and centered in the y dimension, extended the full length of the textured area. (A rectangular white tower 300 ft wide and 1,500 ft tall was generated at the far end of the white stripe to give the pilot a second reference for desired heading. The tower will be deleted in future testing, since it was visible only toward the end of the fastest adaptation events, and the white stripe seemed to be quite adequate by itself.)

In the first experiment reaction times during the pre- and posttest trials were computed from system clock half cycle times to the nearest 1/60 sec. In the second experiment reaction times for all tests, including time to increase rudder-pedal pressure following initiation of a lateral displacement, were recorded in milliseconds on the data tape. Two Durgin and Brown Model LT20BX6B electronic digital timers with BCD output were used for this purpose.

Adaptation design and procedure. Each trial sequence in the experiment consisted of three parts: a pretest, an adaptation period, and a posttest. During the pretests and posttests, the pilot passively observed transformations of the visual scene appropriate to slow forward motion. On half the trials, the motion continued at a constant speed for 6 sec. On the other half, the speed was reduced at a constant rate until the motion stopped. The pilot was instructed as follows: "If the motion

stops, your task is to immediately press the button on the stick that is normally used to control nose-wheel steering.... If the motion continues throughout the test period without stopping, continue to observe the display but do not press the button."

Three initial speeds and deceleration rates were chosen to mimic during the pretest what Denton's[8] observers experienced during the posttest. This strategy was adopted in order to be able to score the pilot's responses as correct or incorrect. The constant speeds were 3.75, 7.5, and 15 kts. The deceleration events had the same initial speeds as the constant, and slowed at 3 kts/sec, stopping in 1.25, 2.5, and 5 sec respectively.

When the pilot responded, or when 6 sec had elapsed if no response was made, the display changed to uniform "gray" for a brief period. (The duration was manually controlled by the console operator during the first preliminary experiment and averaged a little over 4 sec. By the second experiment the interval control was automated and always equalled 3.1 sec.) Following the grayout, the 100-sec adaptation period was initiated with the aircraft centered over the white stripe. Pitch and roll inputs to the visual display were frozen so that the horizon was in a fixed horizontal position throughout the experiment. Altitude was held constant at 100 ft, and speed was constant within a trial at 296.25, 444.35, or 666.50 kts. These speeds produced global optical flow rates of 5, 7.5, and 11.25 eyeheights/sec, respectively. A multiplier of 1.5 was used to determine successive levels of speed.

The aircraft was kept on course directly above the white stripe for 2 sec, at which time the first lateral displacement was induced. The pilot's task was to use only the rudder pedals to get back on course as soon as possible and stay there as long as possible. (The steering task was part of a separate experiment described in detail below.) Left and right displacements simulating 90-deg side wind gusts were repeated throughout the adaptation period. At the end of the period, grayout of the display occurred for 3.1 sec, followed by a posttest. Each postest was identical to some pretest, but never the same test which had just preceded a particular adaptation period.

Thus, the design to test for aftereffects of adaptation was a 3 (Adaptation Speed) by 2 (Pre- versus posttest) by 2 (deceleration versus constant speed Test Type) fully crossed factorial, with three levels of deceleration nested within the deceleration tests. With the constant speeds repeated three times each to match the number of deceleration tests, there were 36 total trial sequences.

The trial sequences were grouped in blocks of three such that (a) each adaptation speed and each initial test speed appeared once in each of the 12 blocks, and (b) deceleration and constant speed tests were used no more than twice in each block, so that no more than four of the same test type could appear in a row.

Order of the three adaptation speeds within a block was determined by using the six possible orders twice. Sequence effects between pretests and adaptation speeds, and between adaptation speed and posttests, were counterbalanced using a series of different 3 x 3 Greco-Latin squares[13]. As a result, each adaptation speed followed each of the six possible pretest conditions an equal number of times, and each adaptation speed preceded each of the six posttest conditions an equal number of times.

Every pilot received three practice trial sequences in the same order. All parameters were the same as those in the experimental trial-sequence blocks, but the specific combinations were different from any other block.

The 12 trial-sequence blocks were then sampled randomly without replacement and presented in a different order for each pilot. Order of the three trial sequences within a given block was the same for every pilot. An example of a block is shown in Table 1.

Time was recorded in milliseconds from onset of the displayed motion until the pilot pushed the button to indicate that the motion had ceased.

Lateral displacement design and procedure. The pilot was started on course directly above the white stripe with altitude frozen at 100 ft and forward speed frozen at one of the three values chosen for the adaptation periods. After 2 sec, a 90-deg crosswind gust was induced

Table 1

An Example of One of the 12 Blocks of Three Trial Sequences

Block Number	Sequence Number	Pretest		Adaptation Period	Initial Lateral Displacement			Posttest	
		$\dot{x}_i$	$\ddot{x}$	$\dot{x}$	k	y_i	Direction	$\dot{x}_i$	$\ddot{x}$
12	34	3.75	-3.00	666.50	20	4.8	L	7.50	0.00
12	35	15.00	0.00	444.35	40	2.4	R	3.75	0.00
12	36	7.50	-3.00	296.25	80	1.2	R	15.00	-3.00

$\dot{x}_i$ = initial speed (in kts) k = fractional loss in speed (in %/sec)

$\ddot{x}$ = deceleration (in kts/sec) y_i = initial lateral displacement (in ft)

$\dot{x}$ = speed (in kts) L = left; R = right

and controlled according to a formula which held the fractional rate of displacement constant:

$$y_t = y_i e^{kT} \qquad 1$$

where y_t = lateral displacement at time t

y_i = initial lateral displacement

e = natural log base

k = constant fractional rate of displacement

Choice of fractional displacement was based on earlier studies of sensitivity to loss in altitude and speed, which indicated that the functional optical variables specify fractional rates of change[3,4].

The pilot was instructed as follows: "Your task is to stay on course as nearly as possible directly above a white stripe perpendicular to the horizon. You will start out on course, but from time to time you will be blown off course by a 90-degree crosswind directly from the right or the left. When this occurs, you are to get back on course as soon as possible. However, you are to use only the rudders to re-establish the correct flight path over the white stripe. This task will be scored in terms of how quickly you get back on course and how much time you are able to stay on course."

To initiate the first displacement, values of initial fractional rate of displacement (k_i) and initial displacement (y_i) were sampled

from among 18 possible combinations. From $\underline{k}_i$ and $\underline{y}_i$, an initial rate of displacement ($\dot{\underline{y}}_i$) was computed according to the formula:

$$\dot{\underline{y}}_i/\underline{y}_i = \underline{k}_i \qquad 2$$

Since values of $\dot{\underline{y}}$ were always positive, both $\underline{y}$ and $\dot{\underline{y}}$ had accelerating functions over time.

In the first preliminary experiment, the values of $\underline{k}$ were 2, 4, and 8%/sec. These proved to be much too low to be easily detected, so the values were changed to 10, 20, and 40%/sec for the second preliminary experiment. Each level of fractional change was produced by three different environmental ratios as shown in Table 2.

Table 2

Environmental Levels of $\dot{\underline{y}}_i/\underline{y}_i$

$\underline{k}_i$(%)	Low	Medium	High
10	0.6/6	1.2/12	2.4/24
20	0.6/3	1.2/6	2.4/12
40	0.6/1.5	1.2/3	2.4/6

If the optical specification of fractional displacement is the only functional source of information for detecting displacement, then performance should improve down the rows of Table 2, but show no change over the columns. If instead or in addition, rate of displacement is useful, performance should improve across the rows. If magnitude of displacement is important, performance will improve along the positive diagonal. (If the environmental variables $\dot{\underline{y}}$ and $\underline{y}$ have significant effects, a test of their global optical specifiers $\dot{\underline{y}}/\underline{z}$ and $\underline{y}/\underline{z}$ would be made by varying, for example, $\dot{\underline{y}}$ and $\underline{z}$ such that their ratio remains the same.)

The nine cells of Table 2 were repeated twice, once for left displacements and once for right, resulting in a 3 (Fractional Rate of Change) by 3 (low, medium, or high Environmental Level) by 2 (left or right Direction) factorial design. The 18 possible initial displacement conditions were assigned to the 12 trial-sequence blocks of the adaptation design so that (a) each fractional rate of displacement and each environmental level was represented in each block, and (b) left and right

displacements were used no more than twice in each block, so that no more than four of the same direction could appear in a row. The order of the conditions within a trial-sequence block was determined by a separate series of 3 x 3 Greco-Latin squares for the nine cells of Table 2. (See Table 1 for an example.) The 18 conditions were used once each in the first six blocks, and repeated once each in the last six blocks.

After the initial computer controlled displacement, the remaining displacements during an adaptation period formed a second experimental design with parameter levels partly dependent on the pilot's behavior. Following an interval of 2 sec from the time that the pilot returned to a zone having a bandwidth of ± 30 ft on either side of the center of the white stripe, the next displacement was induced given that three criteria were met: The center of gravity of the aircraft had to be within the 60-ft zone, and from one cycle to the next (1/30 sec), the aircraft could not have deviated laterally by more than 10 ft or changed heading by more than 5 deg.

Since the pilot would typically have some displacement while within the "on course" criterial envelope, this y_t value was read and used to compute a $\dot{y}_t$ value from the sampled value of k according to Formula 2. The simulated 90-deg side wind was continued until the pilot made an adjustment of either rudder pedal of 0.5 lb or greater from the value 1/30 sec before. At this point, all the environmental variables previously controlled by the computer came under the pilot's control and were recorded as dependent variables until the next displacement was initiated. The six combinations of fractional displacement and direction of displacement were randomly sampled without replacement until all were used, with the constraint that no more than four displacements could occur in the same direction in a row. The randomization procedure was repeated until the end of the adaptation period, then continued into the next period if necessary to finish an incomplete block of six combinations.

Time in milliseconds from onset of the displacement to initiation of the rudder pedal adjustment, was recorded. Performance measures taken every 1/30 sec included recording the rudder pedal chosen,

magnitude of pedal pressure over time, distance and direction of the aircraft from the center line of the white stripe, and fractional change in displacement. Rate of change in displacement was computed from the last two variables. A dichotomous scoring of whether the aircraft was on or off course, by the criteria described previously, was also recorded.

Subjects. Eleven pilots who had just completed a series of training sessions in the F-16 simulator (ASPT) at the Human Resources Laboratory, Williams Air Force Base, Arizona, served as subjects, five in the first preliminary experiment and six in the second. The distribution of total flight time was bimodal, with five pilots having between 220 and 375 hours and six pilots having between 1,000 and 1,600 hours. Four of the latter six were instructor pilots. Two pilots had no low altitude experience, six had between 2 and 20 hours, and three had between 100 and 400 hours. The pilots were current in five different types of aircraft.

Testing of one pilot in the first experiment was terminated after he developed motion sickness at the higher speeds. One other pilot reported nausea halfway through the test session, but continued to the end.

Results. The first preliminary experiment served primarily to debug the testing procedure and to indicate that the fractional rates of displacement were too small. From the posttest results alone, it was concluded that different optical flow rates during adaptation produced different magnitudes of the self-motion aftereffect. The pretest results revealed the same effect, however, where all reaction times should have been the same. This finding demonstrated the value of a pretest-posttest comparison by showing that posttest effects cannot be taken at face value.

The posttest speeds were not great enough to cancel the aftereffect in either experiment. Following the first experiment, it was assumed that the aftereffect was decaying during the 4+ sec taken by the operator to reinitialize the visual system. Shortening and automating this time made it evident that the aftereffect continued to overpower the displayed motion throughout the 6-sec test period. A model of the posttest self-motion experience over time is currently being developed from Denton's[6] unpublished experiments to better predict the speeds needed to cancel the aftereffect within the 6-sec period.

The fractional rates of displacement used in the second experiment appear from initial scanning of a data printout to be nearly optimal. Details concerning transfer of the data to Ohio State are underway, and complete analyses will be carried out in the future.

Results of the pilot post-experimental questionnaire proved very instructive. The pilots were asked whether they felt as though they were sitting still watching a display that represented forward motion, or experienced self motion like that which occurs during actual flight. Five reported sitting still; five reported moving. Estimates of altitude ranged from 10 ft to 1000 ft, including seven different steps between the extremes. Only one pilot estimated the altitude to be 100 ft.

The speeds during the 100-sec adaptation period were estimated relatively accurately by most of the pilots. The pilot who estimated the altitude to be 10 ft, however, judged the speeds to be 30, 60, and 120 knots. Assuming a 10-ft eyeheight, the flow rates were actually appropriate for speeds of 29.6, 44,4, and 66.6 knots. These informal findings suggest that pilots are aware of the eyeheight-flow rate relation.

During the adaptation period, one pilot reported experiencing increasing speed, nine reported constant speed, and four reported decreasing speed. (They were allowed to check more than one alternative.) If adaptation took place at a rapid enough rate to be noticeable, decreasing speed should be reported, otherwise constant speed would be the veridical experience.

Given that pitch and roll were artificially frozen during the displacement tests, the pilots were asked whether the visual transformations following rudder adjustments appeared to be appropriate or inappropriate. Eight pilots reported that the transformations seemed appropriate; two inappropriate. Both voiced essentially the same complaint. One pilot felt that abnormally high force was needed to get the nose tracking, but after that the system was too sensitive to allow smooth adjustments. The other pilot reported that there was little response to slight pressure, and once the larger input required to produce the desired change took effect, it resulted in oscillation back and forth. An attempt will be made to remedy this problem before future testing.

A request for other comments at the end of the questionnaire elicited the following: (a) "Lines in lower area of screen never appeared to stop. When the machine felt like it was stopped, I had the impression I was sinking." (b) "Altitude seemed to vary during the displacement test. Looking toward the horizon gave a sensation of descent to 25-50 ft, while looking right off the nose I felt as though I had climbed to 100-200 ft." (c) "Experienced climbing sensation while looking at the ground close to the nose of the plane." (d) "Motion sickness was worst at high speed, less at medium speed, none at low speed."

Comments (a), (b), and (c) suggest that global specification of optical variables may not be sufficient to account for all the effects of self-motion perception and adaptation. Local effects may differ in magnitude, e.g., greater aftereffects in peripheral vision than central, and different kinds of effects may be experienced depending on where the pilot looks. The second part of Comment (a) suggests an aftereffect of level flight related to perception of change in altitude. This effect is of special interest, since the possibility of such aftereffects was a major motivating force in studying adaptation in the first place.

Comment (d) suggests a relation between flow rate and motion sickness. Drivers in England have reported nausea while driving over the quarter mile of horizontal stripes exponentially decreasing in separation just before a traffic circle. By varying altitude, speed, and distance between horizontal surface texture edges, flow-rate effects can be separated from speed and edge-rate effects using pilots susceptible to motion sickness. Influences of high flow rates and edge rates in producing motion sickness may prove to be important in selecting pilots to fly at low altitude and high speed.

III. RECOMMENDATIONS:

a. Preliminary experimentation has primarily served to work out problems in the testing and data recording procedures and to iterate toward serviceable values of the independent variables. Further work to demonstrate that pilots adapt to simulated flow rates is necessary. Given the competition for ASPT time and the more limited availability of

subjects at the Human Resources Laboratory, it might be more efficient to continue the preliminary experiments in the Aviation Psychology Laboratory at Ohio State University where full-time resources and large numbers of subjects could be devoted to the project over a short period of time. This would (1) have the added benefit of allowing comparison of the two visual simulation systems under conditions as nearly identical as possible, and (2) give a better idea of where pilots selected for low altitude training fall relative to the general population on both adaptation susceptibility and sensitivity to changes in optical flow variables.

b. In an unpublished study, Denton[8] tested observers' time to detect increase or decrease in speed following adaptation periods of 10 sec or 120 sec. Initial speeds ranged from 5 to 80 mph (1.6 to 26 eye-heights/sec), accelerations and decelerations were a constant 10% of the adapting velocity, and the observer was told before each trial whether speed would increase or decrease. Deceleration was noticed more quickly than acceleration. Time to become aware of acceleration decreased with an increase in the duration of the adaptation period, whereas time to notice deceleration increased. If the finding with acceleration generalizes to optical acceleration during descent, it is possible that sensitivity to loss in altitude improves with adaptation, rather than showing a decrement.

Interpretation of Denton's results is not at all straightforward, however, for the following reasons: (1) He had no way to determine the accuracy of his observers, and sensitivity is more appropriately indexed by accuracy than by time to detect. Response times are interpretable only within the context of accuracy, since in different situations the two measures may have a high positive correlation (indicating that both index difficulty of detecting change) or a high negative correlation (indicating a speed/accuracy tradeoff). A forced-choice between "accelerating" and "decelerating" or between either alternative and "constant speed" would allow scoring of the reports. Forced-choice tasks also allow an unbiased estimate of sensitivity, in case one response category is used more frequently than the other. (2) Denton's use of a constant

rate of change in speed resulted in fractional rates of change that differed in both direction and rate. Fractional change for deceleration positively accelerated over time, whereas that for acceleration negatively decelerated. If fractional rate of optical transformation is the functional information for detecting change in speed, as the results of Owen et al.[4] indicate, then Denton's conditions are not at all comparable. Information for detecting deceleration became increasingly salient over time; that for acceleration became decreasingly salient. A fair comparison could be made by holding fractional rates constant throughout each event ($\ddot{x}_t / \dot{x}_t = k_t$) and using the same constant for both acceleration and deceleration trials.

Denton's results are nevertheless intriguing and directly relevant to the problem of change in sensitivity following adaptation. In most of his earlier adaptation experiments, Denton contrasted a group of observers who experienced little or no motion aftereffect with a group showing large aftereffects. He found great differences in both adaptation to self motion stimulation and aftereffects, but tested only the High Motion Aftereffect group in his study of time to detect change following different periods of adaptation. Differences in sensitivity of the two classes of observers, both before and after adaptation, is obviously of interest relative to pilot selection and evaluation. The issues deserve further study with a more objective paradigm and better control of stimulation.

If adaptation effects can be demonstrated in the flight simulation environment, studies of change in sensitivity following adaptation should be pursued. Given continued work on detection of change in speed and altitude at Ohio State and the fact that the ASPT system has been readied for such tests, implementation of this second phase of the research should be possible without encountering too many difficulties. Since the F-16 control stick operates by force alone rather than being moved, it provides an optimal test situation for studying pilot produced visual information because the only feedback from control adjustments is visual.

c. Taking into account the three pilots who were dismissed due to equipment failure (all had 220 hours total flying time), seven of the

14 had 220 hours and six had 1,000 to 1,600 hours. If this sample is even close to being representative of future classes, it should be easy to categorize the pilots into two relatively distinct levels of experience. While there is no particular rationale for relating total hours flight time per se to either adaptation or sensitivity to change, it is possible that selective factors may be at work in determining which pilots achieve more experience, and these factors may very well be related to sensitivity and adaptation to self-motion information. Denton[6,7,8] found great individual differences in adaptation effects and aftereffects and the author and his colleagues[3,4], have found wide variation in sensitivity to optical flow transformations. If possible, it is recommended that pilot experience be included in an equal-N analysis of variance design, rather than used as a covariate, so that interactions with the independent variables of primary interest can be assessed.

d. Preliminary perusal of the data from compensatory adjustments by the pilots in response to the lateral forcing functions ("wind gusts") during the adaptation period suggests that detailed analysis by computer is warranted. As soon as transfer of the data to the Ohio State system is functional, these analyses will begin. In addition to providing the first opportunity for tests of our approach to studying a pilot's perception-action cycle, lateral displacement is a dimension of self motion that we have not previously explored. In addition to the possibility of support for our principle of fractional rates of optical change as functional invariants for self-motion perception, the results have relevance for the practical problem of pilots compensating for strong side-wind gusts during low altitude flight or just before touchdown during landing.

REFERENCES

1. Gibson, J. J. Ecological optics. Vision Research, 1961, 1, 253-262.

2. Gibson, J. J. The ecological approach to visual perception. Boston: Houghton Mifflin, 1979.

3. Owen, D. H., Warren, R., Jensen, R. S., Mangold, S. J., & Hettinger, L. J. Optical information for detecting loss in one's own forward speed. Acta Psychologica, 1981, 48, 203-213.

4. Owen, D. H., Warren, R., & Mangold, S. J. Optical information for detecting loss in one's own altitude. Perception & Psychophysics, in press.

5. Warren, R. Global optical bases for the perception of egospeed. Unpublished manuscript, 1981.

6. Denton, G. G. A study of motion induced percepts of velocity. Unpublished doctoral dissertation, University of Southhampton, 1973.

7. Denton, G. G. The influence of adaptation on subjective velocity for an observer in simulated rectilinear motion. Ergonomics, 1976, 19, 409-430.

8. Denton, G. G. Visual motion aftereffect induced by simulated rectilinear motion. Perception, 1977, 6, 711-718.

9. Gibson, J. J. Visually controlled locomotion and visual orientation in animals. British Journal of Psychology, 1958, 49, 182-194.

10. Powers, W. T. Behavior: The control of perception. Chicago: Aldine, 1973.

11. Owen, D. H., & Warren, R. Perceptually relevant metrics for the margin of aviation safety: A consideration of global optical flow and texture variables. Proceedings of the Conference on Vision as a Factor in Military Aircraft Mishaps, in press.

12. Gum, D. R., Albery, W. B., & Basinger, J. D. Advanced Simulation in Undergraduate Pilot Training: An overview (AFHRL-TR-75-59(I)). Wright-Patterson Air Force Base, OH: Air Force Human Resources Laboratory, Advanced Systems Division, December 1975.

13. Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

THE ROBUSTNESS OF UNIDIMENSIONAL ITEM RESPONSE THEORY MODELS

Prepared by:	Dr. Charles K. Parsons
Academic Rank:	Assistant Professor
Department and University:	College of Management Georgia Institute of Technology
Research Location:	Human Resources Laboratory, Brooks Air Force Base
USAF	Dr. Malcolm Ree
Date:	August 24, 1982
Contract No:	F49620-82-C-0035

ACKNOWLEDGEMENT

The author would like to thank the Air Force System Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to spend an enriching summer at the Human Resources Laboratory at Brooks Air Force Base in San Antonio, Texas.

He would especially like to thank Dr. Malcolm Ree for his assistance and suggestions on this project.

THE ROBUSTNESS OF UNIDIMENSIONAL ITEM RESPONSE THEORY MODELS

by

Charles K. Parsons

ABSTRACT

No set of test data will perfectly fit an item response theory (IRT) model. Test unidimensionality is frequently assumed, yet strict adherence is unlikely in any test. This study examines violations of this assumption and shows that meaningful interpretations can be made when the test deviates from unidimensionality. By varying the degree of correlation among factors, the number of factors, and the distribution of major factor loadings, it was found that more highly correlated factors, more factors, and balanced factor loadings improve the fit of the unidimensional model. Applications to test development and validity are offered.

I. INTRODUCTION

The application of Item Response Theory, IRT,(Lord, 1980) has become a key part of the Air Force personnel testing program. The 3 parameter logistic model (Birnbaum, 1968) is one item response model that is widely studied and used in the program. An important feature of this model as it is currently used is the assumption that the latent trait is unidimensional. On the basis of this assumption, it would appear that increasing fidelity of the items to the unidimensionality assumption would improve the application of the model. However, this perspective may be shortsighted. This paper presents an argument for not using unidimensionality as the sole criterion for test development and then presents data demonstrating that unidimensional IRT models can be applied to multidimensional tests. It also provides some guidelines for developing heterogeneous tests that are amenable to the unidimensional IRT model.

Perhaps the major application of psychological measurement is the selection and placement of individuals into educational or occupational positions. At a basic level, psychological tests provide a sample of behavior. The predictive validity of a test in a particular setting depends, in no small part, on how well this sample of behavior requires mental operations that are fundamental to the behavior required in the particular educational or occupational setting.

A test that is developed to predict performance in applied settings is likely to deviate from the unidimensionality assumption.In applied settings, mental operations are likely to be wide and general in nature. For instance, on a job, a worker may be required to read technical manuals and diagrams, take verbal instruction, etc. in order to repair a machine. A test that considers these many facets of information

input will probably deviate dramatically from the assumption of unidimensionality. On the other hand, a test may contain only items that measure one facet of mechanical reasoning which is likely to yield a more nearly unidimensional test. However, the more valid test would be the former because it requires mental manipulation of a wider range of symbols and information. In other words, in striving to develop tests of high fidelity to the IRT model, one could subvert the original intent of validity.

This paradox need not prohibit us from pursuing both outcomes simultaneously. There is a need to understand how deviations from the unidimensional model affect IRT item and ability parameter estimation and the meaning of these estimates. Unidimensionality can be expressed statistically as

$$P(u_i = 1|\theta) = P(u_i=1|\theta, u_j, u_k, \ldots,) \ (i=k) \qquad (1)$$

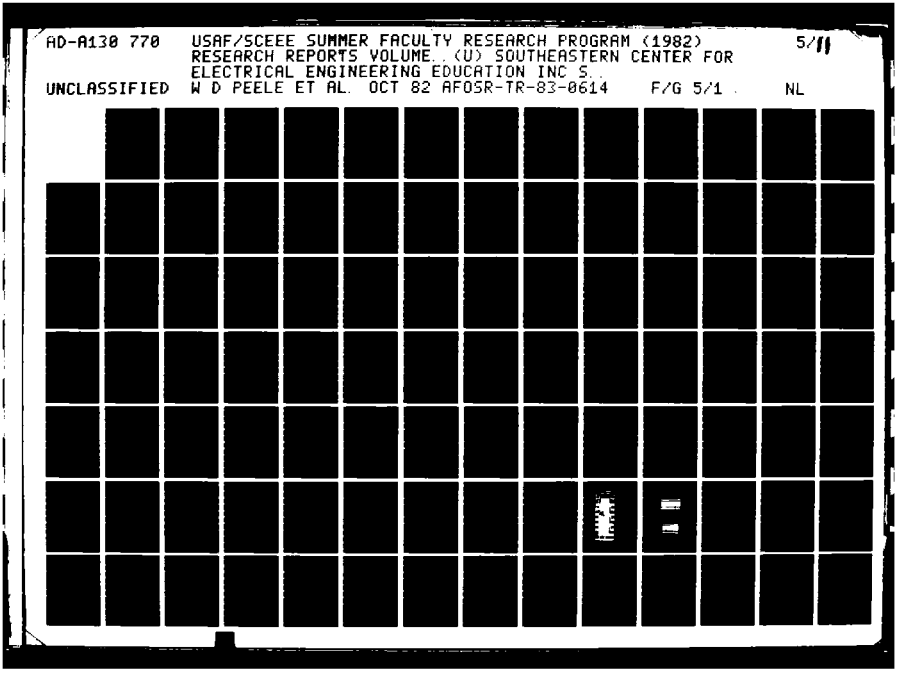
where u_i is an item response scored 1 if correct and 0 otherwise and θ is the latent trait (Lord, 1980; p.19). In other words, once θ is known, there is no other information about the probability of a correct response for a given item in any other item response or responses. Practically, this means that all items in the test measure one and only one ability. For instance, a unidimensional test of mathematical reasoning would contain items for which an examinee's mathematical reasoning ability exclusively determines the probability of a correct response. Other potential factors, such as an examinee's reading ability or test savvy, do not systematically affect responses to two or more items.

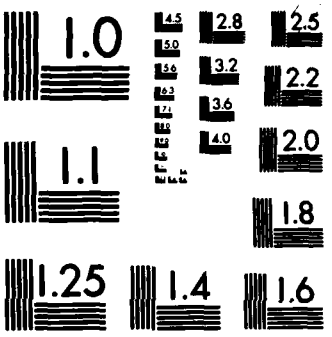
Recently, Drasgow & Parsons (1982) described a Monte Carlo procedure for generating item response data from a known multidimensional factor model. Their results suggested that unidimensional IRT models can be usefully

applied to tests whose underlying factor structure deviates substantially from unidimensionality.

Their basic model of multidimensional tests had a general latent ability such as verbal comprehension, that affected all items on the test and more specific abilities that were each limited to a subset of items on the test. Examples of logical subsets might be analogy type items and sentence completion items. The unidimensionality assumption may not hold strictly in this case because an examinee might consistently perform better on analogies than sentence completion items. The latent trait of interest is general verbal ability and is represented by what these two subsets have in common. By varying the degree of correlation between the abilities underlying subsets of items, the general trait becomes more or less dominant in the total test. The formal model will be described in the next section.

Drasgow and Parsons' results indicated that the general factor could be recovered by a unidimensional IRT estimation procedure, LOGIST (Wood, Wingersky, & Lord, 1976) when factor correlations were as low as .46 to .60. In addition to the degree of factor correlation, other conditions could lead to deviations from unidimensionality. One condition is the number of factors underlying the responses. More factors will lead to greater heterogeneity. A second condition is how the items are distributed across factors. One specific ability might affect more items while the remaining specific abilities affect fewer items. Finally, the factor purity of items within any subset might be less than perfect. Item responses could be slightly, but reliably, affected by other factors besides the general and specific abilities. All of these conditions are likely to occur naturally because no data are fit perfectly by a theoretical IRT model and because the test developer makes choices concerning how many of each item type to


AD-A130 770 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM (1982) RESEARCH REPORTS VOLUME..(U) SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION INC S..
5/11
UNCLASSIFIED W D PEELE ET AL. OCT 82 AFOSR-TR-83-0614 F/G 5/1 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

include in a test. The research presented in this paper is designed to aid the test developer in realizing the implications of such choices on the applicability of the unidimensional IRT model to the resultant test.

II. RESEARCH OBJECTIVES

The goals of the present study were to investigate the effects of factor loading distribution, number of factors, and factor purity on recovery of the general latent trait and item characteristic curves relating to it by the unidimensional IRT model. Two levels of factor intercorrelation were chosen that corresponded to conditions specified by Drasgow and Parsons (1982) to be somewhat critical to the acceptable fit of the unidimensional IRT model. That is, the fit appeared to shift from acceptable to unacceptable in this range.

The IRT model was restricted to the 3 parameter logistic model to maximize applicability of the results to multiple choice tests. This seems justified for the purposes of the Air Force where large samples of examinees are available for parameter estimation.

III. APPROACH

All item response data used in this study were simulated by computer generation. Multidimensional data structure can be represented by the common factor model (Thurstone, 1947) but further transformation is necessary to adequately model dichotomously scored item responses. The common factor model represents continuous observed variables, x, as weighted linear combinations of hypothetical common traits or factors, y, that account for covariation among the variables and unique factors, e, that account for some variance of x but not covariance. This model can be written as

$$x = Ay + Be \qquad (2)$$

where x is a vector containing n observed variables x_i, A is the n-by-k matrix of loadings on the k common factors, y is a vector containing k common factor scores, y_i, B is an n-by-n diagonal matrix with loadings along its diagonal and e is a vector containing the n unique variables e_i. The unique variables are assumed to be mutually uncorrelated and uncorrelated with the common factors. For item response data, the x variables can be thought of as unobservable response propensity variables that underlie the observed dichotomous responses. Let α_{ij} denote the loading of the i^{th} response propensity variable on the j^{th} factor and let β_i denote the single loading of the i^{th} response propensity variable on the i^{th} unique factor.

In the present study, the common factor loading matrices, A, are specified to represent different characteristics of tests. First, major factor loadings are defined as loadings between .4 and .8 in the A matrix. This range was selected to represent what one might expect to find in practice before extensive item screening. Second, minor factor loadings are defined as small, but occasionally non-trivial, loadings that appear as elements in the A matrix that are not specified to be major loadings. These minor loadings represent the "noise" that is inherent in even well developed tests. For the current study, these minor loadings (mean=.000, standard deviation=.098) were randomly selected from the minor loadings in an oblique factor solution (factor correlations ranged from .18 to .32) to an item level factor analysis by Parsons and Hulin (in press).

Four factor loading matrices, A, can be constructed as follows. All matrices represent 50 items or (n=50). The number of factors, k, was set to either 5 or 10. The distribution of major factor loadings was either balanced or unbalanced. In the balanced matrix, each factor had an equal

number of major loadings. In the unbalanced condition, one group factor had 15 major loadings and the remaining major loadings were equally distributed across the remaining k-1 factors. When k=10, 1 factor had 15 major loadings, 8 factors had 4 major loadings, and 1 factor had 3 major loadings. An example of an A matrix with k=5 and unbalanced major loadings is presented in Table 1.

To achieve the loading matrix in Table 1 in aptitude test data, an oblique rotation would probably be necessary. Let Φ represent the matrix of first order factors after rotation to the simple structure in Table 1. It is fundamental to the simulation model presented here that there is one second order factor underlying the matrix of first order common factors. This factor is the general ability factor described earlier.

Schmid and Leiman (1957) describe a hierarchical factor model that is based on a transformation of the k oblique first order factors to a matrix with k+1 orthogonal factors where the additional factor is the general factor represented in terms of item reponse propensity variables rather than first order factors. Basically, the larger the values in Φ, the stronger the general factor. The details of this transformation can be found in Drasgow and Parsons (1982).

For the current study, two levels of factor intercorrelation were chosen. These correspond to factor correlation level 3, FC3 (range of .46 to .60) and factor correlation level 4, FC4 (range of .25 to .39) from Drasgow and Parsons (1982) where higher levels than those used here were clearly fit well by the unidimensional model and lower levels fit poorly. When combined with the number of factors and balanced versus unbalanced major loadings, a 2x2x2 design results.

The k+1 hierarchical factor matrix is used to compute item response

propensity scores, x_i. To do this, a vector of K+2 factor scores is generated to represent the general factor, the k common first order factors, and the e_i (unique) factor. All factor scores were generated as independent, normal (0,1) variables by the Univac Math-Stat package subroutine, RANDN. These factor scores are linearly combined according to eq. (2). The continuous variable, x_i, is transformed to a dichotomous variable, u_i, as follows. First, as part of the simulation model, each item, i, is assigned a value, denoted by γ_i, that is related to the desired proportion of examinees knowing the correct response (p). Then the continuous response propensity variable, x_i, is compared to γ_i and

$$u_i = 1 \text{ if } x_i \geq \gamma_i$$

and (3)

$$u_i = 0 \text{ if } x_i < \gamma_i \ .$$

γ_i equals that point on the abscissa of the normal distribution (mean 0, standard deviation 1) to which the desired proportion of the population knowing the correct response (p) is represented by the area under the curve to the right of γ_i and 1-p lies to the left of γ_i. See Lord & Novick (1968, p. 370) for futher discussion of this transformation. For the current study, the γ_i are presented in Table 1. These values were specified to approximate items where the proportion of examinees knowing the correct response to an item were .1, .2, .3, .4, .5, .6, .7, .8, .9. The values were equally represented among items by randomly selecting without replacement from 50 γ_i values and assigning them sequentially to the 50 items.

Note that γ_i only specifies the proportion of examinees that know the correct response, not those who guess the correct response. In order to simulate items where guessing is possible, the parameter, c, is included

which specifies the probability of an examinee obtaining the correct response to item i, if the examinee does not know the correct response. If $P_i(\theta)$ is the probability of knowing the correct response, then

$$P_i^*(\theta)=c_i+(1 - c_i)\, P_i(\theta). \qquad (4)$$

IV. DATA GENERATION

For any desired factor model, a pseudo examinee's item score vector, U, $(u_1, u_2, \ldots, u_n)$ can be obtained by generating a vector, y, computing $P^*(\theta)$ for all n items, and then generating n random numbers on the uniform interval $[0,1]$. For each item i, $u_i=1$ if $P_i^*(\theta) \geq$ the i^{th} random number and $u_i=0$ if $P_i^*(\theta) <$ the i^{th} random number.

1500 pseudo-examinees were generated for each of the eight factor models described. Each of these 8 samples was input to LOGIST (Wood, Wingersky, & Lord, 1976) to obtain item and ability parameter estimates. These estimates were then compared to the factor models using the following summary statistics.

V. CRITERIA FOR EVALUATION OF PARAMETER ESTIMATES

Lord & Novick (1968, p.375) derive two important relations between the 2 parameter normal ogive model and a factor analysis model with one common factor. First,

$$a_i = \frac{\alpha_i}{\sqrt{1-\alpha^2}} \qquad (5)$$

where a_i is the item discrimination parameter for the i^{th} item and α_i is the loading of x_i on the one common factor. They also showed that

$$b_i = \frac{\gamma_i}{\alpha_i} \qquad (6)$$

where γ_i is the item threshold value defined earlier.

Though mathematical proofs of such relations do not exist for the commonly used logistic IRT models, the normal ogive model and two parameter

logistic model are very similar when the scaling constant, D, is set to 1.7 (difference in $P_i(\theta) < .01$ in the range $-3 \leq \theta \leq +3$).

The relation between factor model and IRT models with guessing have not been derived. However, using the relations specified above, we can expect to approximate whatever true relation does exist. Later in this paper, several empirical comparisons demonstrate this point.

In evaluating the effect of multidimensionality, Drasgow and Parsons (1982) used several summary statistics based on these approximate relations. First, they computed the similarity between LOGIST estimated item parameters, $\hat{a}$ & $\hat{b}$, and the factor model parameters transformed by eqs. 5 & 6 ($\tilde{a}$,$\tilde{b}$). Note that $\underline{\hat{a}}$ and $\underline{\hat{b}}$ are only parameter estimates whereas $\underline{\tilde{a}}$ & $\underline{\tilde{b}}$ are population values. For each data set, the similarity between $\underline{\tilde{a}}$ and $\underline{\hat{a}}$ was assessed by computing the Root Mean Squared Difference, $RMSD_a$, between $\underline{\tilde{a}}$ & $\underline{\hat{a}}$ across the items or

$$RMSD_a = \sum_{i=1}^{n} (\tilde{a}_i - \hat{a}_i)^2/n \qquad (7)$$

where n is the number of items. Likewise, $RMSD_b$ can be computed as

$$RMSD_b = \sum_{i=1}^{n} (\tilde{b}_i - \hat{b}_i)^2/n \qquad (8)$$

In addition, Drasgow and Parsons computed correlations between the LOGIST estimated θ and the <u>general</u> factor scores used to generate the item response data.

In the present study, two additional summary statistics will be computed. Because the latent trait of interest is the general trait, we can infer that we are interested in estimating item characteristic curves that are based on $P(\theta)$ where θ is the general factor score. Therefore, we can compute the estimated ICC's based on item parameter estimates, $\hat{a}$, $\hat{b}$, $\hat{c}$, and compare them to the curves representing the response probabilities based on

the transformed population general factor loadings ($\tilde{a}$), gammas ($\tilde{b}$), and guessing probabilities (c). The differences between the curves can be assessed as an RMSD across 30 equally spaced θ values between -3 and +3 or

$$\text{RMSD(ICC)} = \sum_{j=1}^{30} (\tilde{P}^{*}_{i_j}(\theta) - \hat{P}^{*}_{i_j}(\theta))^2/30 \qquad (9)$$

where $\hat{P}^{*}_{i}(\theta)$ is based on parameter estimates for the i^{th} item and $\tilde{P}^{*}_{i}(\theta)$ is based on the factor weights, threshold values and guessing parameters. Also of concern is the over-estimation or under-estimation of the factor model based ICC by the estimated logistic ICC. Therefore, we also computed BIAS statistic which is simply the algebraic mean of the deviations at the 30 points or

$$\text{BIAS} = \sum_{j=1}^{30} (\tilde{P}^{*}_{i_j}(\theta) - \hat{P}^{*}_{i_j}(\theta))/30. \qquad (10)$$

The advantage of these latter two statistics (RMSD(ICC) and BIAS) is that they effectively summarize differences between ICC's rather than item parameters. Large differences in item parameters are not necessarily related to large differences in ICC's (Hulin, Lissak, & Drasgow, in press; Linn, Levine, Hastings & Wardrop, 1981).

VI. RESULTS

The first results compare normal ogive and 3 parameter logistic item characteristic curves. Four levels of factor loadings, α_i(.5,.6,.7,.8) were randomly paired with 4 levels of γ_i(-.83,-.26,+.26,+.83) and 2 levels of c_i(.15, .20) to obtain 4 ICC's. These values were appropriately transformed using eqs. 5 & 6 to obtain $\underset{\sim}{\tilde{a}}$,$\underset{\sim}{\tilde{b}}$, & $\underset{\sim}{c}$. These latter values were used to compute ICC's according to the 3 parameter logistic model. Differences between ICC's were computed and appear in Table 2. The maximum difference between ICC's is .0094 and the maximum mean absolute difference is .0053 (ICC 2). These differences are well within the tolerance set for the difference between the 2 parameter logistic curve and normal ogive curve (Lord, 1980; p. 14).

Correlations among Item Parameters

The summary statistics for the analysis of the eight data sets generated from the various factor models appear in Table 3. First, note the correlations between LOGIST $\hat{\underline{a}}$ and the $\tilde{\underline{a}}$ (r_a) from the factor models. The effect of factor correlation (FC) is apparent as the lowest correlations appear in 3 of the 4 cells with the lower level of factor correlation (r's = .336, .691, .711). The lone exception occurs when the major loadings are balanced across 5 factors (r = .885). The correlations between $\hat{\underline{b}}$ and $\tilde{\underline{b}}$ (r_b) are higher in general but still lowest in the same 3 cells as the r_a (.907, .693, .922). Again the cell with balanced major loadings on five factors has a high correlation (r_b = .977).

Also, the effect of balanced versus unbalanced major loadings is apparent. Without exception, at the same level of FC and number of factors, the balanced models provided more estimable logistic item parameters. The results are more dramatic for $\underline{a}$ than $\underline{b}$. Probably most interesting are the correlations for FC4. The correlation for a = .885 and for b = .977 when there are balanced loadings on five factors. The correlations are not so impressive for 10 factors. In fact, the effect for number of factors shows an advantage to 5 factors for $\underline{a}$ in all cells and advantages for $\underline{b}$ in 3 of 4 cells.

Item Parameter RMSD

The results for $RMSD_a$ differ somewhat from the item parameter correlations. For the most part, higher factor correlations lead to lower $RMSD_a$. However, in the balanced condition with 5 factors, FC4 has the lower $RMSD_a$ (.237 vs. .255). The balanced models also had consistently lower $RMSD_a$ than their unbalanced counterparts. The largest difference from the parameter correlation results appears when examining the effect of

the number of factors. Ten factors consistently gave lower $RMSD_a$ than 5 factors. In fact, $RMSD_a$ was lower in FC4 with balanced loadings than any condition for 5 factors. The value (.232) was even lower than that reported by Drasgow and Parsons (1982) for 5 factors with even more highly correlated factors than reported here (FC2 in the earlier report).

Concerning $RMSD_b$, the most dramatic effect is for FC in 10 factors. In two conditions, a large number of items have extreme b's and accompanying large sampling errors. Drasgow and Parsons (1982) also noted this tendency in their data.

Correlations between Factor Score and $\hat{\theta}$

For the most part, the correlations between $\hat{\theta}$ and general factor scores follow the pattern for the other statistics. Balanced models are slightly better with an average r ($\bar{r}$) = .801 for balanced and $\bar{r}$ = .771 for unbalanced. There is virtually no effect for the number of factors with $\bar{r}$ = .790 for 10 factors and $\bar{r}$ = .783 for 5 factors. As in the Drasgow and Parsons study, factor correlations had a large effect ($\bar{r}$ = .841 for FC3 and $\bar{r}$ = .731 for FC4). From these results, it appears that some improvement in correlation between factor score and $\hat{\theta}$ is possible by balancing major loadings, but not by increasing the number of specific ability factors underlying the item set.

Item Characteristic Curve RMSD

By computing differences between item characteristic curves, the problems associated with extreme values of b are not as pressing. First, the average BIAS statistic, which would have indicated consistent over or under-estimation by LOGIST of the general trait ICC, were trivial. The values ranged from -.011 (under-estimation) to .001 (over-estimation) which are very similar to the expected differences between the normal ogive and logistic curves.

The average RMSD between ICCs (RMSD(ICC)) do provide some useful differences. First, the average RMSD(ICC) = .050 for all unbalanced models and .039 for balanced models. Second, the average RMSD(ICC) for 10 factor data is .046 and the average RMSD(ICC) for 5 factor data is .044. Third, there is an average RMSD(ICC) of .037 for FC3 and an RMSD(ICC) of .053 for FC4.

In interpreting the results for this study, it is beneficial to consider how much of the dissimilarity is attributable to sampling error and how much is attributable to multidimensionality. Drasgow and Parsons (1982) reported that for unidimensional data with 1500 examinees and 50 items, the $RMSD_a$ was .209, the $RMSD_b$ was .194, and the correlation between factor score and $\hat{\theta}$ was .939. The present results show that when the major loadings are balanced, the multidimensional distortion of $\underline{a}$ is very slight, ranging from .210 for 10 factors at FC3 to .255 at FC3 with 5 factors. The additional "error" is thus between .001 and .046 and quite small relative to the sampling error. The error is much larger when the major loadings are unbalanced. However the correlations between factor score and $\hat{\theta}$ are substantially lower in the present data (maximum correlation = .858).

The RMSD(ICC) results can also be compared to results from a Monte Carlo study by Hulin et al. (in press) that examined the effects of test length and examinee sample size on differences between population and sample ICC's. Because Hulin et al. did not investigate a test of 50 items and sample of 1500 examinees, some linear interpolation was performed to obtain the estimated RMSD(ICC) of .038 in a unidimensional test. From Table 3, it is clear that many of the RMSD(ICC) obtained in the present study were below this value.

VII. DISCUSSION AND RECOMMENDATIONS

The results presented in this study tell an interesting, but somewhat

inconsistent story. Only in the case of differences in factor correlation is there strong agreement among the summary statistics. This is not surprising because of the overall strength of this effect demonstrated in the earlier Drasgow and Parsons (1982) study. It is also quite clear that balancing the major loadings across factors will improve estimation of item parameters and therefore item characteristic curves quite dramatically in multidimensional tests.

The improvement in estimation of the general factor score is not so great. However, the benefit of improved recovery of the general factor in the range of .010 to .085 might be useful in some cases. This improved recovery is still probably insufficient for item pools where factors are correlated only in the range of FC4

Choices about the number of factors presents an interesting issue. One does not expect to introduce additional factors through sampling items from theoretically distinct domains, but rather through sampling items from various related facets of the same general domain. Results that show a better recovery of the general factor through additional specific ability factors may appear illogical at first glance. One reason for this is that if the unidimensionality assumption implies that one factor alone is the best situation, then two factors should be better than three, three better than four, etc. This logical sequence is contradicted by the data.

One explanantion for this paradox is that the factors in the simulation model are not orthogonal, but oblique in a very special way. The factor correlations can be perfectly represented by a single common factor. This provides the general hierarchical factor which appears to be estimated better by IRT methods when more first order factors are inherent in the data. If orthogonal rather than oblique factors had been chosen, the

results could have been much different. In addition, as more factors are added, the effect of any one factor is decreased. Each factor accounts for less common variance. At the extreme, where k (the number of factors) approaches n (the number of items), the common factors become much like item-specific factors. Then the factor model approaches unidimensionality because it only has the one general factor with the multiple item-specific factors.

One can then ask whether or not the results generalize to real data. The input factor model was designed to represent item pools that had not been carefully screened to approach unidimensionality but rather to represent a construct that is somewhat multi-faceted or for which multiple item formats and content are possible. The reported consistent positive correlations among measures of different cognitive abilities (Humphreys, 1962) suggests that orthogonal factors would rarely, if ever, be representative of performance on aptitude tests.

As mentioned in the introduction, an IRT theorist might argue that unidimensional item pools are a necessary condition for proper applications of the unidimensional IRT model. Humphreys (1979) has argued that a test model that leads to highly homogeneous tests actually detracts from the construct validity of the test. High construct validity results from a large collection of items that are "heterogeneous in content and have moderate levels of inter-item correlations" (Humphreys, 1979; p.115) He also notes that a more heterogeneous test is likely to have higher predictive validity for tasks that include a cognitive component.

The important point of the above remarks is that there would appear to be a contradiction in the suggestion to build tests of items with widely varied content which reduces homogeneity and to build tests that are

amenable to unidimensional IRT models and estimation procedures. When the current study's results are added to the earlier Drasgow and Parsons (1982) results, this contradiction is shown to be empirically invalid. The current study also suggests that test developers can probably both improve the construct and predictive validity and still apply IRT methods by tapping many item contents that possible cut across the trait of interest and to avoid their own predispositions to favor one content area over another. This puts an additional burden on any test developer to define the item domain for a construct before writing and testing items.

As for the results showing consistently lower RMSD(ICC) than in previous studies of unidimensional tests, the following should be considered. When a values are very low, large differences in b values have little effect on the ICC. In the current study, the multidimensional factor structures tended to decrease the magnitude of the loadings on the general factor. In the extreme, as these loadings approach zero, there is a "floor" on the a estimate because LOGIST does not allow zero or negative a values. Therefore the possible estimation errors were restricted by the "floor" on the estimated parameters and these errors did not have as great an effect on the ICC as when the a parameter is of higher magnitude. Earlier studies of the unidimensional test simulation (Hulin et al., in press), used a much higher mean and range of a parameters. So the smaller differences in ICCs in the current study appear to be at least partially the effect of average a parameters. Future research could investigate this issue by setting higher limits on the minimum loading on the general factor.

In summary, the current study examined the effect of the number of

factors, balanced versus unbalanced loadings, and factor purity on the application of unidimensional IRT models to multidimensional tests. Results suggest that more factors and balanced loadings lead to improved estimation. Compared to earlier studies, factor impurity did not markedly affect parameter estimation. Finally, these results indicate that a test developer can maintain reasonable IRT estimation while increasing item heterogeneity by tapping many facets of the construct with an equal number of items.

References

1. Birnbaum, A. In F.M. Lord & M.R. Novick, Statistical Theories of Mental Test Scores, Reading, Mass.: Addison Wesley, 1968.

2. Drasgow, F. & Parsons, C.K. Application of unidimensional item response theory models to multidimensional data. Presented at the 1982 Computerized Adaptive Testing Meetings in Minneapolis, MN. (To appear in Proceedings).

3. Hulin, C.L., Lissak, R.J. & Drasgow, F. Recovery of two and three parameter logistic item characteristic curves: A monte carlo study. Applied Psychological Measurement, in press.

4. Humphreys, L.G. The construct of general intelligence. Intelligence, 1979, 3, 105-120.

5. Humphreys, L.G. The organization of human abilities. American Psychologist, 1962, 17, 475-483.

6. Linn, R., Levine, M.V., Hastings, N.C. & Wardrop, Item bias in a test of reading comprehension. Applied Psychological Measurement, 1981, 5, 159-173.

7. Lord, F.M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

8. Lord, F.M. & Novick, M.R. Statistical Theories of Mental Test Scores, Reading, Mass.: Addison Wesley, 1968.

9. Parsons, C.K. & Hulin, C.L. An empirical comparison of item response theory and hierarchical factor analysis in applications to the measurement of job satisfaction. Journal of Applied Psychology, in press.

10. Schmid, J. & Leiman, J. The application of hierarchical factor solutions. Psychometrika, 1957, 22, 53-61.

11. Thurstone, L.L. Multiple Factor Analysis. Chicago: University of Chicago Press, 1947.

12. Wood, R.L., Wingersky, M.S. & Lord, F.M. LOGIST- A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton,N.J.: Educational Testing Service, 1976.

Table 1

Sample Factor Weight Matrix and Threshold Values

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	γ
1	.400	.099	.019	-.024	-.094	.850
2	.500	.031	-.080	.042	.026	- .530
3	.600	.172	.016	-.117	.027	- .850
4	.700	-.065	.137	.129	- .011	- .260
5	.800	-.126	-.130	-.080	- .047	.260
6	.400	.040	-.012	.099	.004	-1.300
7	.500	-.057	-.084	-.070	.073	.530
8	.600	.007	.210	-.156	.070	-.850
9	.700	-.109	-.067	.176	.054	.000
10	.800	.012	.009	.052	-.210	1.300
11	-.002	.400	.023	.027	.007	1.300
12	.029	.500	.197	-.036	.041	-1.300
13	-.022	.600	-.080	-.016	-.076	.530
14	.027	.700	.023	-.028	-.027	-.530
15	.009	.800	.095	.028	.058	.000
16	-.080	.400	.075	-.024	.043	-.850
17	-.027	.500	.061	.038	-.060	-1.300
18	-.064	.600	.022	.231	-.108	.850
19	-.040	.700	.031	-.156	.081	.000
20	.008	.800	.052	-.028	-.034	-.260
21	-.079	.064	.400	.137	-.126	-1.300
22	.195	-.062	.500	-.088	-.067	-1.300
23	.217	-.002	.600	.012	.014	.530
24	-.073	.072	.700	-.029	-.174	.850
25	.047	.090	.800	-.084	.087	.000
26	-.028	-.060	.400	-.065	.110	-.530
27	-.057	.142	.500	.031	.034	1.300
28	.140	.197	.600	.035	-.019	-.850
29	-.030	-.067	.700	.088	-.057	-.260
30	-.019	.120	.800	-.047	-.076	.260
31	-.156	-.016	-.088	.400	.115	-1.300
32	-.133	-.113	-.040	.500	.210	.000
33	-.027	-.023	.115	.600	.020	.850
34	.073	.113	-.012	.700	.009	.530
35	-.131	-.063	-.182	.800	.012	-1.300
36	.117	.012	-.065	.400	-.088	-.850
37	-.070	-.044	.044	.500	-.050	.260
38	.038	.133	-.131	.600	.035	-.260
39	.028	-.077	.075	.700	-.067	-.530
40	.006	-.007	.130	.800	.000	1.300
41	.077	.130	-.024	-.070	.400	.530
42	.119	.095	.073	.070	.500	1.300
43	-.048	-.036	.107	.031	.600	-1.300
44	-.084	.129	.185	-.245	.700	-1.300
45	-.058	.054	.076	-.072	.800	-.260

Table 3 (continued)

46	.028	-.120	-.088	.070	.400	-.530
47	-.156	.175	-.137	-.038	.500	.850
48	-.073	-.267	-.042	.175	.600	.260
49	-.115	.072	.058	-.113	.700	-.850
50	.063	.075	-.143	.031	.800	.000

Note: For all simulation models the c parameter (guessing) was set to .20 for odd numbered items and .15 for even numbered items.

Table 2

Differences between 3 Parameter Logistic ICCs

and Normal Ogive with Guessing ICCs

	Parameters[a] a	b	c	Mean Absolute[b] Difference	Maximum Absolute Difference
ICC1	.4(.579)	.26(.52)	.15(.15)	.0049	.0081
ICC2	.6(.750)	-.83(-1.38)	.15(.15)	.0053	.0088
ICC3	.7(.980)	.83(1.19)	.20(.20)	.0044	.0087
ICC4	.8(1.333)	-.26(- .33)	.20(.20)	.0044	.0094

[a] 3 parameter logistic parameters are in parentheses.

[b] Differences computed at 121 equal θ intervals in the range $-3 \leq \theta \leq +3$.

Table 3

Summary Statistics for 8 Data Sets

		Unbalanced	Balanced
10 Factors	FC3	r_a[1] = .769 r_b = .976 $RMSD_a$ = .214 $RMSD_b$ = .496 r_θ = .848 BIAS = -.011 RMSD(ICC) = .038	r_a = .870 r_b = .973 $RMSD_a$ = .210 $RMSD_b$ = .576 r_θ = .858 BIAS = -.007 RMSD(ICC) = .032
	FC4	r_a = .336 r_b = .907 $RMSD_a$ = .376 $RMSD_b$ = 1.237 r_θ = .684 BIAS = -.003 RMSD(ICC) = .070	r_a = .691 r_b = .693 $RMSD_a$ = .232 $RMSD_b$ = 5.158 r_θ = .769 BIAS = -.006 RMSD(ICC) = .043
5 Factors	FC3	r_a = .889 r_b = .693 $RMSD_a$ = .319 $RMSD_b$ = .652 r_θ = .822 BIAS = .001 RMSD(ICC) = .041	r_a = .900 r_b = .977 $RMSD_a$ = .255 $RMSD_b$ = .538 r_θ = .836 BIAS = -.009 RMSD(ICC) = .036
	FC4	r_a = .711 r_b = .922 $RMSD_a$ = .406 $RMSD_b$ = 1.088 r_θ = .731 BIAS = .000 RMSD(ICC) = .052	r_a = .885 r_b = .977 $RMSD_a$ = .237 $RMSD_b$ = .694 r_θ = .741 BIAS = -.008 RMSD(ICC) = .044

[1] r_a is the correlation between $\tilde{\underline{a}}$ and $\hat{\underline{a}}$.

r_b is the correlation between $\tilde{\underline{b}}$ and $\hat{\underline{b}}$.

$RMSD_a$ is the root mean squared difference between $\tilde{\underline{a}}$ and $\hat{\underline{a}}$.

$RMSD_b$ is the root mean squared difference between $\tilde{\underline{b}}$ and $\hat{\underline{b}}$.

r_θ is the correlation between the general factor score and $\hat{\theta}$.

BIAS is the average algebraic difference between the estimated logistic ICC and the input normal ogive ICC.

RMSD(ICC) is the root mean squared difference between the estimated logistic ICC and the input normal ogive ICC.

1982 USAF-SCEEE SUMMER RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

R-RELIABLE INTERVALS FOR THE SUM OF TWO INDEPENDENT NORMAL OR UNIFORM RANDOM VARIABLES

Prepared by:	Dr. S. A. Patil
Academic Rank:	Professor
Department and University:	Department of Mathematics and Computer Science Tennessee Technological University
Research Location:	Air Force Weapons Laboratory Albuquerque, New Mexico
USAF Researcher:	Chris Ashley
Date:	August 13, 1982
Contract No.:	F-49620-C-82-0035

R-RELIABLE INTERVALS FOR THE SUM OF TWO INDEPENDENT NORMAL OR UNIFORM RANDOM VARIABLES

by

S. A. Patil

ABSTRACT

An R-reliable interval and its associated confidence level C(R) for the sum of two independent random variables is defined. The implications for this confidence level if the additional assumption is added that the summands are normal are explored. This is done for R-reliable intervals based on sample ranges. Also, an R-reliable interval based on the sample variances is investigated, and its confidence level C(R) is tabulated for selected sample sizes and ratios of the standard deviations. It is shown that when the sample sizes of the summand variables are greater than two, the confidence level for an R-reliable interval for the sum is greater than the minimum of the confidences for the corresponding R-reliable intervals for the summand variables. Furthermore, an R-reliable interval and its confidence for the sum of two independent uniform random variables is obtained, and the non-decreasing confidence conjecture of Locasso[5] for the R-reliable interval for that sum is proved also.

Acknowledgements

The author wishes to express his appreciation to the Air Force Office of Scientific Research and to the Southeastern Center for Electrical Engineering Education for the opportunity to perform research on R-reliable intervals at the Air Force Weapons Laboratory, Kirtland AFB, New Mexico. I am grateful for the hospitality which has been generously extended by the laboratory staff.

I wish especially to acknowledge the administrative support of Dr. A. H. Guenther, Chief Scientist, and I'd like to thank Chris Ashley for suggesting this topic for research and for helpful discussions.

My sincere thanks are due to Richard Chavez for his help in computing Table 1.

Preface

Suppose Y is a one-dimensional random variable. Let $Y_1, \ldots, Y_m$ be a random sample. Then we can calculate an exact (i.e., not asymptotic) confidence level for the assertion that $P(Y_{(1)} \leq Y \leq Y_{(m)}) \geq R$. We can do this in at least two different ways (viz., Bayes and Neyman-Pearson) which make no assumptions whatever about the distribution of Y. (That is, either of the two confidence levels is completely nonparametric.) These two confidence levels are:

$$C_y(R) = 1 - R^m \quad \text{(Neyman-Pearson, Wilks)}$$

$$C_y(R) = 1 - R^{m+1} \quad \text{(Bayes)}$$

Suppose Z is a second random variable. Let $Z_1, \ldots, Z_n$ be a random sample. Consider the sum X = Y+Z. What we need is a similar exact nonparametric confidence level for the probability that the sum X is contained in some interval.

Suppose the methods above have already been applied to the variables Y and Z separately to learn that

$$C(P(L_Y \leq Y \leq U_Y) \geq R_Y) = C_Y$$

and

$$C(P(L_Z \leq Z \leq U_Z) \geq R_Z) = C_Z \ .$$

Then it follows fairly easily from the axioms of probability and confidence and the definition of independence that

$$C(P(L_Y+L_Z \leq X \leq U_Y+U_Z) \geq R_Y R_Z) \geq C_Y C_Z \ . \qquad (1)$$

This fact has been known for quite some time; for example, it was stated and used by Ashley[1].

This inequality is completely nonparametric (as well as exact), in the sense that it assumes nothing whatever about the distributions of Y and Z. The reason this totally distribution-free result is not perfectly satisfactory is,

the products $R_Y R_Z$ and $C_Y C_Z$ become small too rapidly for many applications (as the number of factors in the two products increases, for example). So what we really need to do is relax slightly, as little as possible, the requirements for a completely distribution-free result (the foregoing inequality already supplies that), so as to obtain a result which covers <u>almost</u> all (but not <u>all</u>) distributions and yet yields usefully large values of reliability and confidence.

Suppose, for example, that for some distributions of Y and Z the inequality above could be altered by replacing the product $R_Y R_Z$ with the significantly larger value $\min\{R_Y, R_Z\}$. This replacement would considerably improve the utility of that inequality. Suppose further that for some of those distributions it could also be shown rigorously that the product $C_Y C_Z$ could also be replaced with the larger value $\min\{C_Y, C_Z\}$. The inequality would then be much more useful yet (especially if the number of summands was much larger than just two). It would then have become

$$C(P(L_X \leq X \leq U_X) \geq \min\{R_Y, R_Z\}) \geq \min\{C_Y C_Z\} \quad , \qquad (2)$$

for appropriate choices of L_X and U_X. In fact, it has been conjectured (by Locasso[5]) that the second inequality does indeed hold if Y and Z have normal distributions, with $U_X = + (S_Y^2 + S_Z^2)^{1/2}$ and $L_X = -U_X$. In the present paper, we show that, as long as the sample sizes are greater than 2, Locasso's conjecture is true.

It remains to discover for what other distributions than the normal Locasso's conjecture (inequality (2)) is true. In the present paper, we present additional work on this question. In general it would be quite useful to discover necessary and/or sufficient conditions on the distributions of Y and Z for Locasso's conjecture to hold. Of course, though, the replacements with minimums which Locasso has suggested may not be the only way to strengthen

inequality (1) enough to make it really useful.

Therefore the goal of this effort is to discover a way to establish confidence for R-reliable intervals $[L_X, U_X]$ for the sum X at known, exact, essentially nonparametric (i.e., extremely robust), yet usefully large levels.

Because this is an ambitious goal, it is necessary to proceed in stages. In this first publication we accept several limitations. First, for the present we confine our work to the Neyman-Pearson definition of confidence. (We would like to extend it to include Bayesian confidence as soon as possible.) Also, it will be seen that we are approaching "almost all distributions" in a very gingerly manner: we are beginning by examining in this paper the normal and uniform distributions only. We hope to extend the results to other distributions and to families of distributions in the future.

I. INTRODUCTION

Wilks[8] investigated the R-reliable interval* [L,U] with Neyman-Pearson confidence coefficient C such that

$$P\left[\int_{L}^{U} f(x)dx \geq R\right] = C(R) \qquad (1.1)$$

where L,U are functions of a random sample of a random variable X and f(x) is the probability density function (pdf) of X. He obtained a nonparametric R-reliable interval for X. Wilk[7] has given an extensive survey of order statistics and their application to R-reliable interval and has given an extensive list of references on that subject. Guenther[4] discussed the problem of R-reliable intervals for normal Gamma and Poisson distributions. More recently, this subject has also been investigated by Locasso[6] and Bevensee[2]. If L and U are statistics (functions of a random sample of a random variable), then a definition of a Neyman-Pearson R-reliable interval is given by

$$P[P(L \leq X \leq U) \geq R] = C(R) \ . \qquad (1.2)$$

This definition is equivalent to the definition given by Wilks[8]. Suppose Z denotes a random variable which is the (negative of the logarithm of the) stress on a component of a complex system (such as an airplane), and Z denotes the random variable which is the (logarithm of the) threshold of failure of components of that type to that kind of stress. Then the sum X = Y + Z is the safety margin of components of that kind to stress of that type, in bels. One is interested in finding an R-reliable interval on the sum and its associated confidence level, knowing R-reliable intervals for Y and

---

*In his first paper on this subject, Wilks[8] referred to this interval as a "100 R% tolerance interval." Guenther[4] called them "R-content intervals." In the present development, we prefer to follow the more recent notation of Locasso[6] and Bevensee[2,3], and so we refer to it as an R-reliable interval.

Z and their associated confidence levels.

The distributions of Y,Z are not in general known. However, under some circumstances, they might be assumed to be normal. Locasso[6] and Bevensee[2] investigated the consequences of imposing this assumption. Or Y and Z might be assumed to have a bounded variation; for example, they might be assumed to have uniform distributions. In the present paper, the consequences of each of these assumptions is further investigated. Using R-reliable intervals for Y and Z, R-reliable intervals for X and their associated confidence levels $C_x(R)$ are characterized.

II. OBJECTIVES

The purpose of this research is twofold: to define the R-reliable interval and its associated confidence level for the sum of two independent random variables, for example, stress and threshold; and, to study the R-reliable interval and its associated confidence level when the additional assumption is imposed that the underlying summand distributions are either both normal or both uniform.

III. AN R-RELIABLE INTERVAL FOR THE SUM OF TWO INDEPENDENT NORMAL RANDOM VARIABLES WITH KNOWN MEANS

1. Introduction. In this section, we discuss the problem of finding an R-reliable interval for the sum of two independent normal random variables when their means are known, and compare the Neyman-Pearson confidence for this interval with the confidences for R-reliable intervals for the summands. Locasso[5] worked on this same problem. He obtained an expression for the confidence for an R-reliable interval for a sum, and compared this confidence numerically with the confidence levels for the summands when the sample sizes were m = n = 16. He showed that the confidence level

associated with the sum is no less than the confidence level associated with each component. (The summand confidence levels were equal because the sample sizes were equal.) We shall discuss the problem of finding an R-reliable interval using sample ranges, and compare the confidence levels.

2. The R-reliable interval for the sum. Since we assume that the means of the normal random variables are known, wolog (without loss of generality) the means can be taken to be zero. Let $Y_1,Y_2,\ldots,Y_m$ be a random sample from $N(0,\sigma_Y^2)$, $Z_1,Z_2,\ldots,Z_n$ a random sample from $N(0,\sigma_Z^2)$, and $\sigma_Z^2 = A^2\sigma_Y^2$. We assume wolog that $A \geq 1$. (If $A < 1$, we can interchange Y and Z and thus force $A \geq 1$.)

Also, we define

$$L_X = L(Y_1,Y_2,\ldots,Y_m,Z_1,Z_2,\ldots,Z_n)$$

and

$$U_X = U(Y_1,Y_2,\ldots,Y_m,Z_1,Z_2,\ldots,Z_n)$$

and specify R. Thus, the R-reliable interval for $X = Y + Z$ is determined by

$$P[P(L_X \leq X \leq U_X) \geq R] = C_X \,. \qquad (3.1)$$

Here, C_X on the right-hand side is the confidence level of the R-reliable interval $[L_X,U_X]$ for X. We consider

$$L_X = -(W_Y^2 + W_Z^2)^{1/2}, \qquad U_X = (W_Y^2 + W_Z^2)^{1/2},$$

where W_Y and W_Z are sample semi-ranges of the Y and Z samples. These statistics are suggested by Locasso[6]. Then (3.1) can be written as

$$P[P(-(W_Y^2 + W_Z^2)^{1/2} \leq X \leq (W_Y^2 + W_Z^2)^{1/2}) \geq R] = C_X \,. \qquad (3.2)$$

Since Y is distributed as $N(0,\sigma_Y^2)$, Z is distributed as $N(0,\sigma_Z^2)$, and Y and Z are independent, it follows that X is

distributed as $N(0,\sigma_Y^2 + \sigma_Z^2)$. Using that fact, we write

$$C_X = P[P(\frac{X}{\sigma_Y(1+A^2)^{1/2}} \leq (\frac{W_Y^2 + W_Z^2}{\sigma_Y^2(1+A^2)})^{1/2} \geq R]$$

$$= 1 - F_{(W_Y^2+W_Z^2)/\sigma_Y^2}[z^2_{(1+R)/2}(1 + A^2)] \quad , \quad (3.3)$$

where $z_{(1+R)/2}$ is the $(1+R)/2^{th}$ quantile of a standard normal random variable. For $R = .9$, $z_{(1+R)/2} \doteq 1.645$. $F_{(W_Y^2+W_Z^2)/\sigma_Y^2}(u^2)$ denotes the cdf (cumulative distribution function) of $(W_Y^2+W_Z^2)/\sigma_Y^2$. We note that the distribution of $(W_Y^2+W_Z^2)/\sigma_Y^2$ does not depend on σ_Y^2, but involves A. The determination of C_X requires the distribution of $W_Y^2/\sigma_Y^2 + A^2W_Z^2/\sigma_Z^2$. Let $f_m(w_Y)$ and $f_n(w_Z)$ denote the pdf's of W_Y/σ_Y and W_Z/σ_Z, respectively. Then the distribution function in (3.3) can be expressed by

$$P[(W_Y^2 + W_Z^2)/\sigma_Y^2 \leq z^2_{(1+R)/2}(1 + A^2)] = \iint f_m(w_Y)f_n(w_Z)dw_Y dw_Z$$

$$0 \leq W_Y^2 + A^2W_Z^2 \leq u^2(1 + A^2) \quad (3.4)$$

where u is defined to be $z_{(1+R)/2}$. The region of integration is bounded by an ellipse in the first quadrant, and is shown in Figure 1.

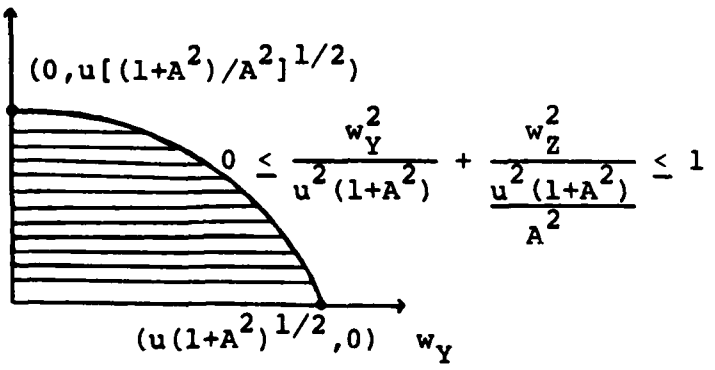


Figure 1

Integrating on w_Z first, C_X reduces to

$$C_X = 1 - \int_0^{u(1+A^2)^{1/2}} F_n\left[\left(\frac{u^2(1+A^2)-w_Y^2}{A^2}\right)^{1/2}\right] f_m(w_Y)\,dw_Y \qquad (3.5)$$

where $F_n[\cdot]$ denotes the cdf of W_Z/σ_Z. This distribution cannot be put into closed form, as it involves the integral of a standard normal random variable. However, $F_n(w)$ and $f_m(w)$ are both tabulated for m = 1 to 16, and n = 1 to 16, by Harter[5]. These can be used to find C_X by numerical integration.

3. Comparison with individual confidence levels. An R-reliable interval and its associated confidence level for the summand Y is determined by

$$\begin{aligned} C_Y &= P[P(-W_Y \le Y \le W_Y) \ge R] \\ &= P[P(|Y/\sigma_Y| \le W_Y/\sigma_Y) \ge R] \\ &= 1 - F_{W_Y/\sigma_Y}[z_{(1+R)/2}] \ . \end{aligned} \qquad (3.6)$$

Similarly, an R-reliable confidence interval and its associated confidence level for Z is

$$C_Z = P[P(-W_Z \le Z \le W_Z) \ge R] = 1 - F_{W_Y/\sigma_Y}[z_{(1+R)/2}] \ . \qquad (3.7)$$

C_Y and C_Z depend on the sample sizes m and n only through the distribution of a standarized variate, w/σ.

It can be said that the R-reliable interval for X = Y+Z is as good as the R-reliable intervals for Y and Z for a given R:

$$C_X \ge \min\{C_Y, C_Z\} \text{ for all } A \ge 1 \ . \qquad (3.8)$$

Locasso[6] conjectures that (3.8) holds if Y and Z are normal. He verifies (3.8) numerically for m = n = 16, $C_Y = C_Z$, and

$C_X \geq C_Z = C_Y$. For the verification of the conjecture, a set of sufficient conditions is given by

(i) $\lim_{A \to \infty} C_X(A) = C_Z$

(ii) For all $m,n \geq 2$, $C_X(A)$ is a decreasing function of A.

We verify the sufficiency of these conditions using the following lemmas.

Lemma 1. For fixed m and n, $C_X(A)$ approaches C_Z as A approaches infinity.

Proof: Using (3.5), and taking the limit as $A \to \infty$, we find

$$\lim_{A \to \infty} C_X(A) = 1 - \lim_{A \to \infty} \int_0^{u(1+A^2)^{1/2}} F_n\left[\left(\frac{u^2(1+A^2) - w_Y^2}{A^2}\right)^{1/2}\right] \cdot f_m(w_Y)\,dw_Y .$$

Since $\lim_{A \to \infty} F_n\left[\frac{(1+A^2)u^2 - w_Y^2}{A^2}\right]^{1/2} = F_n(u)$, we find

$$\lim_{A \to \infty} C_X(A) = 1 - \int_0^{\infty} F_n(u) f_m(w_Y)\,d\,w_Y = 1 - F_n(u) = C_Z, \quad (3.9)$$

which proves Lemma 1.

Lemma 2. For all $m,n > 2$, $C_X(A)$ is a non-decreasing function of A.

It is not easy to prove Lemma 2. $C_X(A)$ involves the integral whose integrand consists of the distribution function and the pdf of the sample range, neither of which can be put into closed form; thus, to verify the lemma, we use numerical integration of

$$C_X = 1 - \int_0^{u(1+A^2)^{1/2}} f_n\left[\left(\frac{u^2(1+A^2) - w_Y^2}{A^2}\right)^{1/2}\right] f_m(w_Y)\,dw_Y .$$

The functions $F_n(u)$ and $f_m(u)$ are tabulated by Harter[5]; these tabulations could be used in numerical integration of the above integral for m = 2 to 16 and n = 2 to 16, and the lemma can thus be verified numerically. In this way, the Locasso conjecture can be verified.

IV. AN R-RELIABLE INTERVAL BASED ON SAMPLE VARIANCES

1. Introduction. Bevensee[2,3] considered R-reliable intervals based on the sample variances when the population variance σ is treated as random. The natural estimates of population variances are sample variances; therefore, an R-reliable interval and its confidence can be obtained by using sample variances.

2. An R-reliable Interval for the Sum Using Sample Variances. An R-reliable interval for the sum random variable X, based on sample variances, is to be constructed when the population variances are unknown parameters. For this purpose, let $Y_1, Y_2, \ldots, Y_m$ and $Z_1, Z_2, \ldots, Z_n$ be random samples from $N(0,\sigma_Y^2)$ and $N(0,\sigma_Z^2)$, respectively; and let

$$S_Y^2 = \frac{1}{m} \sum_{i=1}^{m} Y_i^2 , \qquad S_Z^2 = \frac{1}{n} \sum_{i=1}^{n} Z_i^2$$

be the sample variances. Since the population means are zero, $V_Y = mS_Y^2/\sigma_Y^2$ has a chi-square distribution with m degrees of freedom and $V_Z = nS_Z^2/\sigma_Z^2$ has a chi-square distribution with n degrees of freedom. The R-reliable interval based on S_Y^2 and S_Z^2 is determined by

$$\begin{aligned} C_X &= P[P(-(S_Y^2 + S_Z^2)^{1/2} \leq Y + Z \leq (S_Y^2 + S_Z^2)^{1/2}) \geq R] \\ &= P[(S_Y^2 + S_Z^2)^{1/2}/\sigma_Y \geq z_{(1+R)/2}(1 + A^2)^{1/2}] \\ &= 1 - P[V_Y + A^2 V_Z \leq z^2_{(1+R)/2}(1 + A^2)] \end{aligned} \tag{4.1}$$

where $A = \sigma_Z/\sigma_Y$, $z(1 + R)/2$ is the $(1 + R)/2^{th}$ quantile of the standard normal distribution. The confidence in (4.1) is obtained by integrating the density function over the region shown in Figure 2.

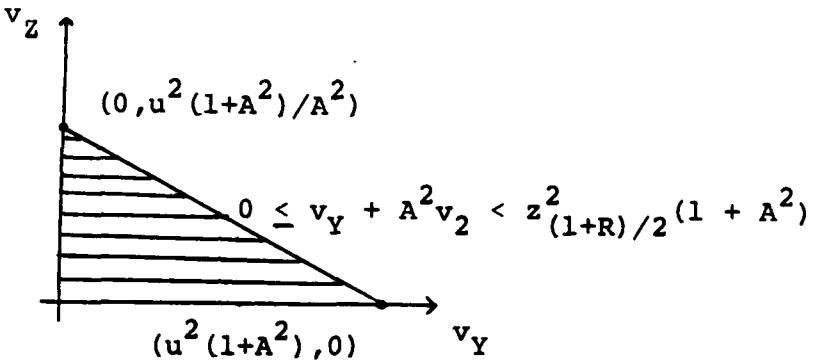


Figure 2

Abbreviating $z_{(1+R)/2}$ by u, and integrating on V_Z first, C_X reduces to

$$C_X = 1 - \int_0^{(1+A^2)u^2} G_n\left(\frac{u^2(1 + A^2) - t}{2A^2}\right) g_m(t)\,dt \qquad (4.2)$$

where $G_n(t)$ and $g_m(t)$ are the distribution function and the pdf of the chi-square distribution with n and m degrees of freedom, respectively. If n or m is even, then the expression for C_X can be further simplified. If $n = 2n_0$, and n_0 is any positive integer, then

$$G_n(u) = 1 - \sum_{j=0}^{n_0-1} e^{-(u/2)} \frac{u^j}{2^j j!},$$

and

$$C_X = 1 - G_m[u^2(1 + A^2)] + e^{-u^2(1+A^2)/2A^2} \sum_{j=0}^{n_0-1} \frac{1}{2^j j! A^{2j}} \sum_{k=0}^{j} \binom{j}{k}$$

$$\cdot (-1)^k [u^2(1 + A^2)]^{j-k} \left(\frac{A^2}{A^2 - 1}\right)^{k+\frac{m}{2}} \frac{2^k \Gamma(\frac{m}{2} + k)}{\Gamma(\frac{m}{2})} G_{m+2k}\left[u^2\left(\frac{A^4 - 1}{A^2}\right)\right]. \tag{4.3}$$

Expression (4.3) can be evaluated for particular values of n, m, and A. Table I gives values of C_X for m = 2, 4, 6, 8, 10, 12, n = 2, 4, and A = 1, 2, 4, 10, 30, ∞.

TABLE I

n=2 A \ m=	2	4	6	8	10	12
1	.2476	.4925	.7127	.8620	.9428	.9792
2	.2453	.3152	.4204	.5289	.6400	.7428
4	.2533	.2702	.2882	.3074	.3280	.3498
10	.2576	.2612	.2628	.2654	.2682	.2708
30	.2582	.2583	.2589	.2592	.2595	.2598
∞	.2585	.2585	.2585	.2585	.2585	.2585

n=4 A \ m=	2	4	6	8	10	12
1	.4925	.7127	.8620	.9428	.9792	.9440
2	.5795	.6646	.7461	.8184	.8779	.9230
4	.6006	.6227	.6450	.6675	.6901	.7128
10	.6069	.6104	.6139	.6175	.6210	.6245
30	.6080	.6084	.6088	.6092	.6096	.6100
∞	.6082	.6082	.6082	.6082	.6082	.6082

3. A Comparison of the Confidence for the Sum with the Confidence for the Summands. The confidence level for the R-reliable interval for Y in terms of S_Y^2 is given by

$$C_Y = P[P(|Y| < S) > R]$$

$$= 1 - G_m(z^2_{(1+R)/2}) . \tag{4.4}$$

Similarly,

$$C_Z = 1 - G_n(z^2_{(1+R)/2}) . \tag{4.5}$$

To compare C_X with C_Y, C_Z, we first show that

$$\lim_{A \to \infty} C_X = C_Z . \tag{4.6}$$

To show this from (4.3), taking the limit as $A \to \infty$ yields

$$\lim_{A \to \infty} C_X = 1 - 1 + e^{-u^2/2} \sum_{j=0}^{n_1-1} \frac{1}{2^j j!} u^{2j} \quad (1)$$

$$= e^{-u^2/2} \sum_{j=0}^{n_1-1} \frac{u^{2j}}{2^j j!} = 1 - G_n(u^2) = C_Z \ . \quad (4.7)$$

From Table 1, it is clear that $C_X \geq \min(C_Y, C_Z)$, for n or $m > 2$ and all $A \geq 1$. However, for $n = m = 2$, the confidence C_X is given by

$$C_X = e^{-u^2(1+A^2)/2} + e^{-u^2(1+A^2)} 2A^2 \left(\frac{A^2}{A^2 - 1}\right)\left(1 - e^{-u^2(A^4-1)/2A^2}\right). \quad (4.8)$$

For $A = 2$, from (4.8) and $u \doteq 1.645$, we find that

$$C_X(2) = e^{-u^2/2} \ (.9492) \ ,$$

and as $A \to \infty$

$$C_X(\infty) = e^{-u^2/2} = C_Z = C_Y \ .$$

Hence, $C_X(2) < C_Z = C_Y$, and Locasso's conjecture fails if (but only if) the sample sizes are both 2.

V. AN R-RELIABLE INTERVAL FOR THE SUM OF TWO UNIFORMLY DISTRIBUTED INDEPENDENT RANDOM VARIABLES

1. Introduction. In sections III and IV, an R-reliable interval for the sum of normal random variables was considered. From the expressions for the R-reliable interval for the sum, it is clear that the confidence level associated with an R-reliable interval can be improved if supplementary information about the distribution of the summands is available. However, distribution-free confidence levels for R-reliable intervals can be obtained for the summand variables by using order statistics. This leads to the consideration

of an R-reliable interval for the sum of uniform random variables. In this section, the R-reliable interval for the sum of uniform random variables with different scale parameters is considered.

2. The R-reliable Interval for the Sum. Suppose Y is distributed uniformly from $-\zeta_Y$ to ζ_Y and Z is distributed uniformly from $-\zeta_Z$ to ζ_Z. Let $Y_1, Y_2, \ldots, Y_m$ be a random sample from $U(-\zeta_Y, \zeta_Y)$ and let $Z_1, Z_2, \ldots, Z_n$ be a random sample from $U(-\zeta_Z, \zeta_Z)$. A one-sided R-reliable interval for the sum $X = Y + Z$ is investigated assuming Y and Z are independent. The confidence level associated with an R-reliable interval for the sum based on the largest absolute value order statistics $|Y|_{(m)}$, $|Z|_{(n)}$ is given by

$$C_X = P[P(-(|Y|_{(m)} + |Z|_{(n)}) \leq Y + Z \leq |Y|_{(m)} + |Z|_{(n)}) \geq R]$$

$$= P[P(0 \leq |Y+Z| \leq |Y|_{(m)} + |Z|_{(n)}) \geq R]$$

$$= P[P(0 \leq Y/\zeta_Y + AZ/\zeta_Z \leq Y_{(m)}/\zeta_Y + AZ_{(n)}/\zeta_Z) \geq R] \quad (5.1)$$

where $A = \zeta_Z/\zeta_Y$. Now $Y:U(-\zeta_Y, \zeta_Y)$ implies $|Y|:U(0, \zeta_Y)$, and $Z:U(-\zeta_Z, \zeta_Z)$ implies $|Z|:U(0, \zeta_Z)$. So from this point on, we will just drop the absolute value signs. Using the distributions of $Y/\zeta_Y + AZ/\zeta_Z$ and $Y_{(m)}/\zeta_Y + AZ_{(n)}/\zeta_Z$, C_X can be simplified to

$$C_X = 1 - mA^{-n}\beta(m,n+1)(t_R)^{m+n}B_{1/uR}(m,n+1) \quad (5.2)$$

where

$$t_R = (1 + A) - [2A(1 - R)]^{1/2}, \text{ for } A \geq 1/2(1 + R) \quad (5.3)$$

$$C_X = 1 - (t_R - A)^n + mA^{-n}\beta(m,n+1)t_R^{m+n}[B_{1/t_R}(m,n+1) - B_{(1-A/t_R)}(m,n+1)] \quad (5.4)$$

where $t_R = AR + .5$ for $A < 1/2(1 + R)$, (5.5)

where $\beta(a,b)$ denotes the complete beta function and $B_x(a,b)$ denotes the ratio of the imcomplete beta function to the complete beta function.

3. Comparison with the Summand Associated Confidence Levels. An R-reliable interval for Y and its associated Neyman-Pearson confidence level using distributions of Y/ζ_Y and $Y_{(m)}/\zeta_Y$, can be written as

$$C_Y = P[P(0 \leq Y \leq Y_{(m)}) \geq R] = 1 - R^m . \qquad (5.6)$$

Similarly, an R-reliable interval and its associated Neyman-Pearson confidence level for Z is given by

$$C_Z = P[P(0 \leq Z \leq Z_{(n)}) \geq R] = 1 - R^n . \qquad (5.7)$$

These facts are proved by Wilks[8] to compare the confidences; the following lemma is used.

Lemma. $\lim_{A \to \infty} C_X(A) = C_Z$.

Proof: Since $A \to \infty$ and $A > 1/2(1-R)$, from (5.4) and (5.5) C_X can be written as

$$C_X = 1 - \left(\frac{AR + .5}{A}\right)^n m \int_0^1 v^{m-1}\left(1 - \frac{v}{AR + .5}\right)^n dv .$$

Taking the limit as $A \to \infty$,

$$\lim_{A \to \infty} C_X = 1 - R^n m \int_0^1 v^{m-1}\, dv = 1 - R^n = C_Z .$$

Next, we prove Locasso's conjecture:

$$C_X \geq \min\{C_Y, C_Z\} \text{ for all } A > 1 \text{ and all } m,n \geq 2 . \qquad (5.8)$$

Two cases must be considered: $A < 1/2(1-R)$ and $A \geq 1/2(1-R)$.

Case 1: $A \geq 1/2(1-R)$. For this $t_R = AR + .5$, and

$$C_X = 1 - m \int_0^1 v^{m-1} \left(\frac{AR + .5 - v}{A}\right)^n dv \ .$$

Without loss of generality, we can take $m \leq n$. Since $R \leq 1$, and $\min\{C_Y, C_Z\} = 1 - R^m$, (5.8) reduces to

$$C_X \geq 1 - R^m \ . \tag{5.9}$$

To verify (5.9), one needs to show

$$J_m(A) \triangleq m \int_0^1 v^{m-1} \left(\frac{.9A + .5 - v}{A}\right)^m dv \leq (.9)^m \ . \tag{5.10}$$

Integrating the middle of (5.10), one finds

$$J_m(A) = \frac{1}{A^m} \sum_{j=0}^{m} \binom{m}{j} (-1)^j (.9A + .5)^{m-j} \frac{m}{m + j} \ . \tag{5.11}$$

It can be seen that $J_m(A)$ is an increasing function of A; hence, $J_m(A)$ approaches maximum when $A \to \infty$. However, from the lemma, the limit of $J_m(A)$ as $A \to \infty$ is $(.9)^m$. Hence, $J_m(A) = \leq (.9)^m = C_Y$, which implies (5.8).

Case 2: If $A \leq 1/2(1-R)$, then $t_R = A + 1 - [2A(1-R)]^{1/2}$, and $R = .9$. Using $n = m$ and (5.5) for Locasso's conjecture, we need to show that

$$C_X = 1 - (1 - \sqrt{.2A})^m - m \int_{1-\sqrt{.2A}}^1 v^{m-1} \left(\frac{A + 1 - \sqrt{.2A} - v}{A}\right)^m dv \geq 1 - R^m$$

or

$$K_m(A) = m \int_{1-\sqrt{.2A}}^1 v^{m-1} \left(\frac{A + 1 - \sqrt{.2A} - v}{A}\right)^m dv \leq (.9)^m - (1 - \sqrt{.2A})^m \ . \tag{5.12}$$

Since $\left(1 + \frac{1 - \sqrt{.2A} - v}{A}\right)$ is a decreasing function of A for $1 - \sqrt{.2A} < v < (2 - \sqrt{.2A})/2$, $1 \leq A \leq 5$, and $\left(1 + \frac{1 - \sqrt{.2A} - v}{A}\right)$

is an increasing function of A for $(2 - \sqrt{.2A})/2 \leq v < 1$, $1 \leq A \leq 5$,

$$K_m(A) \leq m \int_{1 - \sqrt{.2A}}^{(2-\sqrt{.2A})/2} v^{m-1}(2 - \sqrt{.2} - v)^m dv + m \int_{(2-\sqrt{.2A})/2}^{1} v^{m-1}(1 - \frac{v}{5})^m dv = g(A) \ . \quad (5.13)$$

It can be shown that $g'(A) < 0$; hence, $g(A)$ is decreasing in A, and $g(A)$ has maximum at $A = 1$. Integrating $g(1)$, we find

$$K_m(A) \leq g(A) \leq g(1) = m\beta(m,m+1)[d^{2m}(B_{.5}(m,m+1) - B_{(d-1)/d}(m,m+1)) + (5)^m(B_{.2}(m,m+1) - B_{d/10}(m,m+1))] \ , \quad (5.14)$$

where $d = 2 - \sqrt{.2}$.

Using published tables for the incomplete beta function, it can be shown that

$$K_m(A) \leq g(1) \leq (.9)^m - (1 - \sqrt{.2})^m \leq (.9)^m - (1 - \sqrt{.2A})^m, \text{ for } m = 2,3,... \quad (5.15)$$

From (5.15), it follows that $C_X \geq C_Y$ for all $1 \leq A \leq 5$, $m = 2,3,....$ Hence, Locasso's conjecture holds.

VI. CONCLUSION

In this development, R-reliable intervals for the sum of two independent normal random variables, based on sample ranges and sample variances, are obtained. Also, for two independent uniform random variables, an R-reliable interval for the sum and the associated confidence level for this interval are found. These confidences are compared with the corresponding confidence levels for R-reliable intervals for the summands when the summand intervals are assigned the same reliability as the sum. It is shown that when sample

sizes of summands are both greater than two the confidence for the sum interval is at least as great as the smaller of the two confidences for the summands. The result is Locasso's conjecture. It is also shown that Locasso's conjecture fails when the populations are normal and the sample sizes equal two.

It is demonstrated that the confidence level associated with an R-reliable interval for the sum can be improved if the distributions of the summands are known. Therefore, further investigation into other distributional families is desirable. In particular, one is interested in knowing whether Locasso's conjecture would hold when the summand random variables have long-tailed distributions such as the Cauchy distribution.

REFERENCES

1. Ashley, C. (1979), Probabilistic and Statistical Considerations in Assuring Hardness of Complex Systems by Individual Control of Many Independent Piece-parts, DNA System EMP Hardening Symposium held at Naval Ocean Systems Center, San Diego, California, August 1979.

2. Bevensee, R. M. (1980), Proof of a Generalized Locasso Conjecture, Lawrence Livermore National Laboratory.

3. Bevensee, R. M. (1982), A Reliability-Confidence Methodology for Complex Systems, Lawrence Livermore National Laboratory.

4. Guenther, William G. (1972), Tolerance Intervals for Univariate Distributions, Naval Research Logistics Quarterly, 19, 309-333.

5. Harter, H. L. (1969), Order Statistics and Their Use in Testing and Estimation, Vol. 1. Aerospace Research Laboratories, Office of Aerospace Research USAF.

6. Locasso, James V. (1977), The Confidence in Combinations of Imperfectly Known Variances, AFWL EMP Probability and Statistics Note 9, Air Force Weapons Laboratory.

7. Wilks, S. S. (1948), Order Statistics, Bull. Am. Math. Soc., Vol. 54, pp. 5-50.

8. Wilks, S. S. (1942), Statistical Prediction with Special Reference to the Problems of Tolerance Limits, Ann. Math. Stat., Vol. 13, pp. 400-409.

1982 USAF - SCEEE SUMMER FACULTY RESEARCH PROGRAM

Conducted by the

Southeastern Center for Electrical Engineering Education

Sponsored by the

Air Force Office of Scientific Research

FINAL REPORT

MODEL REDUCTION OF CONTROL SYSTEMS

Prepared by: L.R. Pujara

Academic Rank: Associate Professor

Department & University: Engineering & Mathematics Dept.,
Wilberforce University

Research Location: Flight Dynamics Laboratory
Wright-Patterson AFB

USAF Research Colleague: David Moorehouse

Date: October 1982

Contract Number: F49620-82-C-0035

TABLE OF CONTENTS

SECTION		PAGE
I	INTRODUCTION	1
II	FORMULATION OF THE PROBLEM	1
III	DERIVATION OF THE DESIGN ALGORITHM	2
IV	DISCUSSION OF THE ALGORITHM	2
V	A NUMERICAL EXAMPLE AND COMPARISON OF RESULTS	2
VI	SUMMARY	3
	REFERENCES	4
	FIGURE	5

MODEL REDUCTION OF CONTROL SYSTEMS

L. R. PUJARA

ABSTRACT

In this report a computer-aided method of model reduction of single-input/single-output control systems is presented. The simplification is carried out by minimizing a weighted mean square error between the frequency responses of the given and the reduced systems. A numerical example related to the aircraft pitch rate/pilot input of the AFTI/F-16 aircraft is simplified using the proposed technique, and the results are compared with those obtained by the McFit technique.

ACKNOWLEDGMENTS

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and Southeastern Center for Electrical Engineering Education for providing him with this opportunity of creatively interacting with the Air Force Flight Dynamics Laboratory, Wright-Patterson AFB, Ohio, under the Summer Faculty Research Program. He would particularly like to thank Mr David Moorhouse, Group Leader, Flying Qualities Group, for having provided him with excellent working conditions.

The author also wishes to thank Mr David Moorhouse, Mr Frank George and Mr Tom Gentry for several useful, informative and stimulating conversations.

I. INTRODUCTION

The analysis and design of large order control systems is quite tedious and costly. It is, therefore, desirable to replace a given large order system by a lower order system in such a way that the lower order system retains the significant characteristics of the given system.

From the flying qualities standpoint, the usefulness of the model reduction technique lies in verifying compliance with the specifications. The analytical description of a typical augmented aircraft system is usually quite complicated. It, thus, becomes quite difficult to check if the designed aircraft system meets the flying qualities specifications. In order to do this, a reduced order model of a suitable order, depending on the aircraft response being checked, is obtained. The flying qualities requirements are then compared against the parameters of this reduced order model.

During the past decade, several papers have appeared to tackle the model reduction problem. Some model reduction methods use the time-domain techniques while others are based on frequency-domain techniques. The methods of Davison (1966), Anderson (1967) and Aoki (1968) use the time-domain techniques while those of Chen and Shieh (1968), Hsia (1972) and Rao and Lamba (1974) employ frequency-domain techniques. The technique proposed in this report is in the frequency domain and is a modification of the technique of Rao and Lamba (1974).

The technique of Rao and Lamba (1974) is quite effective, but the way it is set up, it cannot be applied to those transfer functions which have either a zero or a pole at the origin. The transfer function to be reduced that was given to the author by the Flying Qualities Group, Flight Dynamics Laboratory, WPAFB, Ohio, had a zero at the origin. This necessitated the modification of the Rao and Lamba (1974) technique. The modified technique as proposed in this report is quite general and is applicable to any transfer function whether it has a zero or a pole at the origin.

Under the proposed technique, the parameters of the lower order transfer function are determined by minimizing a weighted mean square error between the frequency responses of the given and the reduced systems. The weighting function, which linearizes the problem, is the product of the denominators of the given and the reduced systems. A similar weighting function concept was effectively utilized by Levy (1959) in his work on complex curve fitting and by Rattan and Yeh (1978) in their work on digitizing continuous control systems.

II. FORMULATION OF THE PROBLEM

The transfer function, $F(s)$, of a given single variable continuous control system with known real constant coefficients is to be approximated by a lower order transfer function, $G(s)$, with unknown real coefficients to be determined. It is required to determine the unknown coefficients so that the frequency responses of $F(s)$ and $G(s)$ are as "close as possible".

III. DERIVATION OF THE DESIGN ALGORITHM

The frequency responses of the given and the reduced control system are utilized in an error function of the frequency responses. This error function is defined as the integral over the frequency interval of interest, to , of the square of the absolute value of the difference between the frequency responses of the given and reduced systems with respect to the frequency . By equating to zero the partial derivatives of the error function with respect to the unknown coefficients, nonlinear algebraic equations involving the unknown coefficients are obtained. These nonlinear equations can be quite hard to solve and can take a lot of computer time. Now a suitable modification of the error function is made in such a way that the accuracy of approximation is retained, and linear equations involving the unknown coefficients are obtained. This modification is to use the product of the denominators of the given and the reduced transfer functions as a weighting function within the error function. Experience of working with this 'linearizing technique' has shown that this tremendously simplifies the problem and gives fairly accurate results.

IV. DISCUSSION OF THE ALGORITHM

The model reduction technique developed in this report, which is a modification of the Rao and Lamba (1974) technique, is computer-aided; and a Fortran program was developed to implement it. To use the program, the user has to provide the coefficients of the given transfer function, the orders of the numerator and the denominator of the reduced model and the frequency interval (usually determined by the bandwidth considerations) over which the error function has to be minimized. The program yields the optimal parameters of the reduced system as well as the frequency responses of the given and the reduced transfer functions. If the analyst is not satisfied with the response of the reduced model, an increase in the order of the reduced system or adjustment in the frequency interval over which the error is being minimized, will usually get a better approximation. From the distribution of the poles and zeros of the given system, one can predict the order of the reduced system which will give fairly good results. Since the proposed technique is not iterative, it is quite fast, and this offers tremendous flexibility in applications.

V. A NUMERICAL EXAMPLE AND COMPARISON OF RESULTS

In this section a numerical example related to the aircraft pitch rate/pilot input of the AFTI/F-16 aircraft is simplified using the proposed technique. The transfer function considered is a simplification of the complete AFTI/F-16 flight control block diagram in the original q-reconfiguration mode. The transfer function relating the aircraft pitch rate and the pilot input is 8th/13th order. Using the proposed technique over a frequency range of .3 to 10. rad/sec yields a reduced order model of one zero and three poles of

$$G_1(s) = \frac{0.6418s+1.654}{0.01764s^3+0.2103s^2+0.601s+1}$$

Also, a reduced order model of one zero and two poles is obtained yielding

$$G_2(s) = \frac{0.3098s+1.693}{0.1484s^2+0.415s+1}$$

Using the McFit program of McDonnell Douglas based on the search technique of Rosenbrock, the reduced order model with one zero, two poles and a delay element is given by

$$G_3(s) = \frac{e^{-0.0736S}(2.31598608s+20.6998258)}{s^2+4.168408575s+11.147204}$$

Since a delay factor can be approximated by a pole, the reduced model obtained by the proposed technique and given by the 1st/3rd reduced order model would be comparable to the one obtained by the McFit technique and given by the 1st/2nd-plus-delay reduced order model. Figure 1 gives the magnitude plots of the given and the reduced systems. From the flying qualities point of view, the frequency range of interest for this example is approximately .3 to 10 rad/sec. It is clear from Figure 1 that the proposed technique gives a fairly good fit compared to the one given by the McFit method. The main advantage of the proposed technique over the McFit program is that it provides significant savings in computer time. It also gives unique, and therefore repeatable, results.

VI. SUMMARY

In this report a computer-aided method of simplifying single variable control systems is developed. As a matter of fact, it is a modification of the technique of Rao and Lamba (1974). This modification was necessitated by the fact that the Rao and Lamba method cannot be applied to control systems with either a pole or zero at the origin.

The McFit model reduction technique being used in industry as well as WPAFB is iterative in nature and can take a large number of iterations before a solution begins to converge. The proposed method is a one-step procedure and thus provides significant savings in computer time. It also gives unique, and therefore repeatable, results.

REFERENCES

1. G. S. G. Beveridge and R. S. Schecter, "Optimization Theory and Practice".

2. C. F. Chen and L. S. Shieh, "A Novel Approach to Linear Model Simplifications", Int. J. Contr., Vol 8, pp 561-570, 1968.

3. E. J. Davison, "A Method for Simplifying Linear Dynamic Systems", IEEE Trans. Automatic Control, Vol AC-11, pp 93-101, 1966.

4. R. Genesio and M. Milanese, "A Note on the Derivation and Use of Reduced Order Model", IEEE Trans. Automatic Control, Vol AC-21, pp 118-122, 1976.

5. R. Kalman, Trans ASME, Vol 80, pp 468-478, 1958.

6. V. Krishnamurthy and V. Seshadri, "Model Reduction Using the Routh Stability Criterion", IEEE Trans. Automatic Control, Vol AC-23, pp 729-730, 1978.

7. E. C. Levy, IRE Trans. Automatic Control, Vol AC-4, pp 37-43, 1959.

8. S. V. Rao and S. S. Lamba, "A New Frequency Domain Technique for the Simplification of Linear Dynamic Systems", Int. J. Contr., Vol 20, pp 71-79, 1974.

9. A. S. S. R. Reddy, Int. J. Contr., Vol 23, pp 403-409, 1976.

10. D. A. Wilson, "Optimal Solution of Model Reduction Problem," Proc. IEEE, Vol 117, pp 1161-1165, 1970.

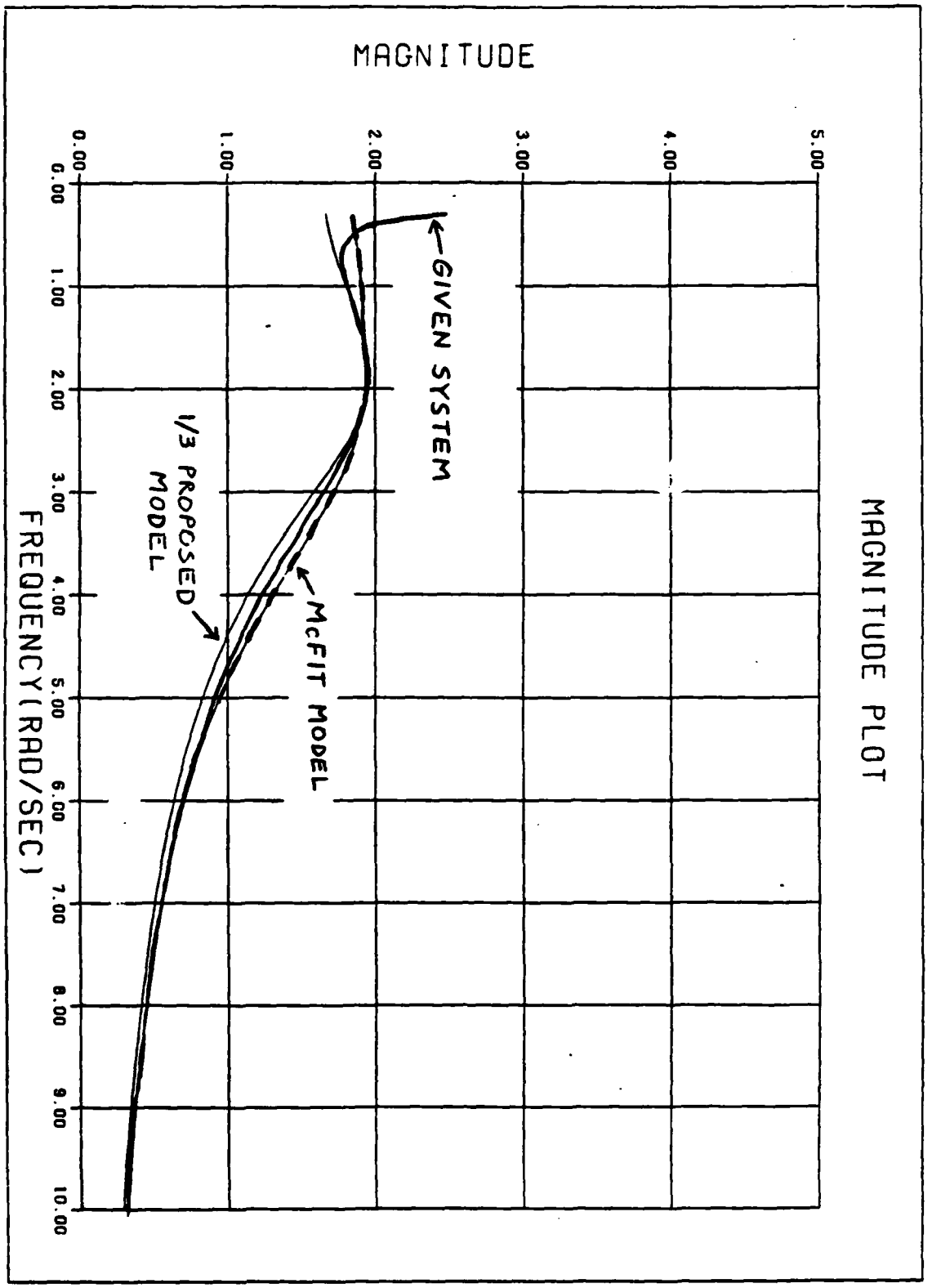


Figure 1. Comparison of Magnitude Plots of Given System, McFit Model and 1st/3rd Proposed Model

1982 USAF - SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

MEASUREMENT OF LEVELS OF PROSTAGLANDIN-E AND PROSTAGLANDIN-F$_{2\alpha}$ IN VARIOUS TISSUES, URINE, AND PLASMA OF RATS GIVEN NONADECAFLUORODECANOIC ACID

Prepared by:	Dr. Zahir Qureshi
Academic rank:	Assistant Professor
Department and University:	Division of Science and Mathematics Education Rust College
Research Location:	Air Force Aerospace Medical Research Laboratory Toxic Hazards Division Toxicology Branch Wright-Patterson AFB, OH 45433
USAF Research Colleague:	Dr. William C. Keller
Date:	September 8, 1982
Contract No:	F49620-82-C-0035

MEASUREMENT OF LEVELS OF PROSTAGLANDIN-E AND PROSTAGLANDIN-$F_{2\alpha}$ IN VARIOUS TISSUES, URINE, AND PLASMA OF RATS GIVEN NONADECAFLUORODECANOIC ACID

by

Zahir Qureshi

ABSTRACT

The effect of 50 mg/kg nonadecafluorodecanoic acid on the levels of prostaglandin-E_2 and prostaglandin-$F_{2\alpha}$ in heart, kidney, liver, urine, and plasma of rats was investigated. Control rats were given 2 ml/kg propylene glycol/water (1:1) solvent. Difficulties in the initial extraction procedure were encountered and, as a result of several discussions with prostaglandin researchers, the extraction procedure was modified. The prostaglandin-E and $F_{2\alpha}$ levels were measured using modified (^3H) prostaglandin radioimmunoassay kits. In the final assay, standard curves could not be drawn due to unreliable counts per minute. A wide variation in the counts per minute for samples of the same tissue is also seen suggesting weaknesses in the technique or malfunction in the scintillation counter. Comparisons between counts per minute for treated and control groups are made but no conclusion is reached. Recommendations are made for further study.

ACKNOWLEDGEMENTS

The support and assistance of the Air Force Systems Command, the Air Force Office of Scientific Research, and the Southern Center for Electrical Engineering Education in providing the author with the opportunity to spend a satisfactory and meaningful summer at the Air Force Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio, is acknowledged with sincere thanks. The author acknowledges the Toxicology Branch for its excellent laboratory facilities and the staff for their hospitality.

The author extends his thanks to Lt Col C. T. Olson for his collaboration and help during the tenure of this research effort. He would like to acknowledge helpful dialogues with Ms. M. E. George and Capt William C. Keller.

I. INTRODUCTION

Perfluorinated compounds are generally considered to be non-reactive in most biochemical systems.[1,2] However, there are published reports which suggest that these compounds have inhibitory effects on enzyme systems.[3,4,5] Some authors have suggested that these compounds stimulate microsomal membrane peroxidation.[6,7] Lipid peroxidation, which is the oxidative deterioration of polyunsaturated lipids, has been reported as a mechanism of cellular injury in some disease processes and in some chemically induced toxicities.[8,9,10,11]

Although the commercial use of these compounds is widespread, very little was known about their metabolism or toxicity until 1976.[12] Since that time, several research findings have been published on the metabolism and toxicity of perfluorooctanoic acid (PFOA).[13,14,15] Toxicity studies on nonadecafluorodecanoic acid (NDFDA), performed at the Air Force Aerospace Medical Research Laboratory, showed a decrease in fetal weights and some maternal and fetal deaths when injected into pregnant rats.[16] Hypocellularity of bone marrow, bile duct proliferation, and hepatocellular alterations in both nucleus and cytoplasm, testicular atrophy and necrosis, mild gastritis with hyperkeratosis of the squamous portion of the stomach, and severe thymic atrophy are commonly observed after exposure to NDFDA. Survival time at LD_{50}, toxic signs and histopathology found in animals injected with NDFDA were similar to those noted in animals injected with 2,3,7,8-tetrachlorodibenzodioxin (TCDD).[17,18] A more recent and detailed study[19] shows a significant shift in fatty acid distribution in the tissues of rats treated with NDFDA. In the

liver there is a relative increase in palmitic and oleic acids and a relative decrease in stearic and arachidonic acids. Similar changes, though lower in degree, are noted in testes and blood.

The mechanism of toxicity of NDFDA, which is a ten-carbon straight chain perfluorinated compound with the formula $CF_3(CF_2)_8COOH$ is not known with certainty. It has been postulated that autoxidation of polyunsaturated fatty acids result in compounds like prostaglandins and leukotrienes which may adversely affect physiological activities.[20,21]

II. OBJECTIVE

The main objective of this study was to measure the levels of prostaglandin-E and prostaglandin-$F_2\alpha$ in heart, kidney, liver, plasma, and urine of rats after the administration of 50 mg/kg of NDFDA. The information obtained from the study is regarded as a necesary component of a larger and more comprehensive study being done at the Air Force Aerospace Medical Research Laboratory, Wright-Patterson AFB to describe a mechanism of action of NDFDA.

III. METHOD AND MATERIALS

Fischer 344[a] rats, obtained approximately two weeks before the onset of experimentation, were maintained at the Air Force Aerospace Medical Research Laboratory's vivarium. Two days prior to their use, the rats were transported to the Toxic Hazards Division laboratory where they were randomly caged individually in a portable containment system[b]. Rat chow[c] and tap water were given ad libitum.

---

a Charles River Breeding Laboratories Inc., Wilmington MA

b PCS-80, Hazleton Systems Inc., Aberdeen MD

c Purina Rat Chow, Ralston Purina Co., St. Louis MI

NDFDA solution was prepared with propylene glycol/water (1:1) as solvent. Injected volumes were 2.0 ml/kg, giving a dose of 50 mg/kg body weight for each rat. All injections were administered intraperitoneally. Additional rats were given 2.0 ml/kg propylene glycol/water (1:1) and served as controls. Eighteen hours prior to sacrifice, rats were placed in metabolic cages to collect urine. On the fourth day after injection animals were anesthesized with halothane in a bell jar and blood samples were drawn in heparinized syringes from either the inferior vena cava or by cardiac puncture. Liver, heart, and kidney were removed immediately and placed in ice-cold methanol to prevent further synthesis of prostaglandins.

1. Prostaglandin-E[d]

Blood plasma: Whole heparinized blood samples were centrifuged at 3,000 rpm and 5° C to separate plasma, from which 2.0 ml were transferred to conical glass centrifuge tubes. To remove neutral fats, 2.0 ml of petroleum ether was added to the tubes, mixed, and the top layer, containing dissolved neutral fats, was removed and discarded. To the remaining phase 2.0 ml of ice-cold methanol was added to stop production of prostaglandins. Then 2.0 ml of ethyl acetate followed by 3.0 ml of distilled water were added, mixed, and the phases were separated by centrifugation. The uppermost organic phase was transferred to polypropylene test tubes and dried at 50° C under air in a hood.

Urine: Urine was collected over a period of approximately 18 to 20 hours. To stop prostaglandin synthesis, 2.0 ml of ice-cold methanol

---

d Modified procedure from "(^{3}H) Prostaglandin-E Radioimmunoassay Kit", Clinical Assays, Division of Travenol Inc., Cambridge MA

was added to 2.0 ml of urine. Two ml of ethyl acetate followed by 3.0 ml of distilled water were added, mixed, and phases separated by centrifugation. The uppermost organic phase was transferred to polypropylene test tubes and dried at 50° C under air in a hood.

Tissue: Approximately one half gram of the tissue was homogenized[e] in 2.0 ml of ice-cold methanol. The homogenate was vortexed for 15 seconds, twice. Then 4.0 ml of ethyl acetate and 3.0 ml of water were added and mixed thoroughly using a vortex mixer. The uppermost organic phase was separated by centrifugation and dried at 50° C under air in a hood.

To convert prostaglandin-E to prostaglandin-B, the dried samples of plasma, urine, and tissues were dissolved in 2.0 ml of gel tris buffer and 1.0 ml of this solution was transferred to two-ml screw-cap vials and 0.1 ml of 1N sodium hydroxide was added. The pH was adjusted to between 12.5 and 12.9 using microelectrodes. The vials were capped tightly, mixed well and placed in a boiling water bath for 5 minutes. The samples were allowed to reach room temperature and the pH adjusted to 7.4 using glacial acetic acid and sodium hydroxide. The fluid volume in each vial was recorded.

The assay procedure included the preparation of a standard curve from which the unknown prostaglandin content in the extracted sample was to be determined. Polypropylene tubes were labelled in duplicate to indicate total counts, dilutions from PGB_1 standard of 40 ng/ml (1/1 with 2,000 pg/0.05 ml/tube, 1/3 with 667 pg/0.05 ml/tube, 1/9 with 222 pg/0.05 ml/tube, 1/27 with 74 pg/0.05 ml/tube, 1/81 with 25 pg/0.05 ml/tube and 1/243 with 8.2 pg/0.05 ml/tube) and samples of plasma, urine, liver,

---

e Tekmar SDT Series Tissumizer, Tekmar Co., Cincinnati, OH.

heart, and kidney from experimental and control rats. One ml of gel tris buffer was added to dilutions from PGB_1 standard and 0.6 ml transferred to all sample extract tubes. Fifty µl of each PGB_1 standard was added to labeled tubes followed by 0.4 ml of extract samples to appropriate tubes. Fifty µl of (^3H) PGB_1 tracer[f] was added to all tubes. Fifty µl of rabbit anti-PGB_1 serum was added to all tubes except total counts and were mixed thoroughly. All tubes were then incubated for 60 minutes at 37° C. This was followed by an addition of 100 µl of normal rabbit serum[f] and 100 µl of goat anti-rabbit serum[f] to all tubes except total counts and mixed thoroughly on a vortex mixer. All tubes were then incubated for approximately 60 hours at 5° C. All tubes except total counts were centrifuged[g] for 1 hour at 3,500 rpm and 5° C. The supernatant was carefully and completely removed and discarded and 1.0 ml of 1N sodium hydroxide was added to all tubes including total counts, and vortexed thoroughly to dissolve the pellet. The solutions were then transferred into scintillation vials and 11.0 ml of scintillation fluid[h] was added and mixed thoroughly by vortexing. All tubes were counted[i] for 5 minutes.

---

f (^3H) Prostaglandin-E Radioimmunoassay Kit; Catalogue No: CA-501. Clinical Assays, Division of Travenol Laboratories, Inc., Cambridge MA.

g Beckman TF-6R Centrifuge, Beckman Instruments, Inc., Palo Alto, CA.

h Scintisol, Isolab, Inc., Akron, OH.

i Packard Liquid Scintillation System TRI-CARB 2660, Palo Alto, CA

2. Prostaglandin-F2α[j]

All steps were the same as for prostaglandin-E except no PGB was used. Instead, the six standards were made from a standard prostaglandin-$F_{2\alpha}$ of 24 ng/ml (1/1 with 2,400 pg/0.1 ml/tube, 1/3 with 800 pg/0.1 ml/tube, 1/9 with 267 pg/0.1 ml/tube, 1/27 with 88.9 pg/0.1 ml/tube, 1/81 with 29.6 pg/0.1 ml/tube, and 1/243 with 9.9 pg/0.1 mg/tube) and (^{3}H) prostaglandin-$F_{2\alpha}$ tracer[k] was used in place of PGB.

The following assay procedure was used. 0.5 ml of gel tris buffer was added to standard tubes and 0.2 ml to sample extract tubes. 100 µl of reconstituted normal-rabbit serum was added to all tubes except total counts. It was followed by 100 µl of goat anti-rabbit serum. The reactants were mixed thoroughly on a vortex mixer. All tubes were then capped and incubated for a minimum of 48 hours at 4^{o} C in a refrigerator. All tubes were centrifuged (RCF 1,600 x g) at 5^{o} C for 60 minutes. The supernatant from all tubes, except total counts, was poured out and tubes carefully cleaned with filter paper to remove all supernatant. This was followed by the addition of 1.0 ml of 0.1N sodium hydroxide to all tubes and the pellet dissolved by mixing using a vortex mixer. The solution was then decanted into scintillation vials, 11.0 ml of scintillation fluid added and mixed and the vials counted for 5 minutes.

---

j Modified procedure from "(^{3}H) Prostaglandin-$F_{2\alpha}$ Radioimmunoassay Kit", Clinical Assays, Division of Travenol Laboratories, Inc., Cambridge, MA.

k (^{3}H) Prostaglandin-$F_{2\alpha}$ Radioimmunoassay Kit, ibid.

IV. RESULTS AND DISCUSSION

Using commercially avialable kits[f,k] a preliminary study to measure prostaglandin-E and prostaglandin-$F_{2\alpha}$ in rat tissues and plasma was attempted in the early part of the summer of 1982. All steps, as described in the kits, were followed carefully. However, there were problems encountered in the first extraction steps. The supernatant, in a substantial number of samples, did not separate after homogenization of the tissue in the solution mentioned in the kits. Instead, a thick gel was formed. Filtration under suction was tried, but did not solve the problem.

Discussion of extraction problems with experts in the field of prostaglandin metabolism led to the modified procedure which has been described in section III, Methods and Materials.

The radioactivity data (counts per minute) for dilutions of standard are given in Table 1. Standard curves for both prostaglandin-E

Table 1: Counts per minute of dilutions of standard for prostaglandins E and $F_{2\alpha}$.

Dilution	1/1	1/3	1/9	1/27	1/81	1/243	1/1	1/3	1/9	1/27	1/81	1/243
Avg cpm	441	215	473	180	240	179	437	142	198	285	311	309

and prostaglandin-$F_{2\alpha}$ cannot be drawn from these data. It appears that one or more steps of the technique were not followed carefully. In the absence of standard curves, the best approach to analyze the obtained data is to adjust the counts per minute to per gram of

tissues and to per ml of plasma and urine, and then make comparisons between the data obtained for NDFDA treated and control rats. This was done and the results are presented in Table 2 and Table 3. An

Table 2: Average prostaglandin-E counts per minute per ml of plasma and urine, and per gram of tissue of experimental (NDFDA) and control (propylene glycol/water, 1:1) rats.

Rat	Experimental Group					Control Group				
No.	Plasma	Urine	Liver	Kidney	Heart	Plasma	Urine	Liver	Kidney	Heart
1	405	588	1011	189	774	331	226	1542	689	597
2	327	278	1321	803	479	241	172	1390	918	594
3	367	381	870	415	558	266	381	470	803	845
4	235	355	479	754	520	191	278	358	662	699
5	219	547	489	604	604	159	275	609	852	661
6	292	327	509	1129	1168	224	281	855	798	533
Range	186	310	842	689	689	172	209	1184	256	312
Mean	308	413	780	684	684	235	269	871	787	655

observation of the ranges shows wide variations in counts per minute for each tissue and urine and plasma, in both experimental and control groups. This variation in data does not lend reliability to any conclusions that may be drawn. The study, at best may be regarded as exploratory.

Table 3: Average prostaglandin-$F_{2\alpha}$ counts per minute per ml of plasma and urine, and per gram of tissue of experimental (NDFDA) and control (propylene glycol/water, 1:1) rats.

	Experimental Group					Control Group				
Rat No.	Plasma	Urine	Liver	Kidney	Heart	Plasma	Urine	Liver	Kidney	Heart
1	10459	835	41423	36672	26820	11395	938	46427	27067	43486
2	9660	693	34039	36206	37704	10861	904	41759	37954	35906
3	10179	1224	40191	32791	32475	10040	938	44937	37113	34358
4	9273	699	33139	35215	41511	9853	1731	40867	30300	34658
5	9273	849	36032	42963	30109	10936	955	34318	31013	32904
6	10070	Lost	41014	29704	47899	10061	744	37752	26303	28218
Range	1186	375	8284	13259	21079	1542	827	12009	11651	15268
Mean	9819	860	37640	35592	36086	10524	1035	41029	31625	34922

V. RECOMMENDATIONS

Since this study did not resolve the question asked in the objective, it is recommended that in future studies in this area:

1. single isotope methods be used to measure the levels of prostaglandins.
2. snap-frozen organs be thawed and homogenized into a denaturing solvent, e.g. methanol, to prevent prostaglandin synthesis prior to extraction.
3. C18 columns be used for extraction prior to radioimmunoassay.

REFERENCES

1. Clark, L. C., Jr., Becattini, F., Kaplan, S., Obrock, V., Cohen, D. and Bechker, C. 1973. "Perfluorocarbons Having a Short Dwell Time in the Liver", Science, 181:680-682.

2. Holaday, D. A. 1970. "Distribution of Fluorocarbon Liquids". Fed. Proc., 29:1815-1816.

3. Triner, L., Verosky, M., Habit, D. V., and Nahas, G. G. 1970. "Perfusion of Isolated Liver with Fluorocarbon Emulsions", Fed. Proc., 293:1778-1781.

4. Spitzer, H. L., Sachs, G., and Clark, L. C., Jr. 1970. Fluorocarbon Effects on Tissue Metabolism, Fed. Proc., 29:1746-1750.

5. Danis, M., Kaufman, F. C., Evans, R. K., and Thurman, R. G. 1981. "Role of Reducing Equivalents from Fatty Acid Oxidation in Mixed-Function Oxidation: Studies with 2-Bromooctanoate in the Perfused Rat Liver", J. Pharmacol. Exp. Ther., 219:383-388.

6. Rozner, M. A. and Taves, D. R. 1980. "Comparison of Lipid Peroxidation by Perfluoro-octanoic Acid or CCl_4", presented at the Society of Toxicology Meeting in March.

7. Ullrich, V. and Diehl, H. 1970. "Uncoupling of Monooxygenation and Electron Transport by Fluorocarbons in Liver Microsomes", Eur. J. Biochem., 20:509-512.

8. Recknagel, R. O. and Ghoshal, A. K. 1966. "Lipoperoxidation as a Vector in Carbon Tetrachloride Hepatotoxicity", Lab. Invest., 15:132-145.

9. Diluzio, N. R. and Hartmen, A. D. 1967. "Role of Lipid Peroxidation in the Pathogenesis of the Ethanol-Induced Fatty Liver", Fed. Proc., 26:1436-1442.

10. Plaa, G. L. and Witson, H. 1976. "Chemicals, Drugs, and Lipid Peroxidation", Ann. Rev. Pharmacol, Toxicol., 16:125-141.

11. Stacey, N. H. and Klassen, C. D. 1981. "Inhibition of Lipid Peroxidation Without Prevention of Cellular Injury in Isolated Rat Hepatocytes", Toxicol. Appl. Pharmacol., 58:8-18.

12. Guy, W. S., Taves, D. R. and Brye, W. S., Jr. 1976. "Organic Fluorocompounds in Human Plasma: Prevalence and Characterization", in Biochemistry Involving Carbon-Fluorine Bonds, Ed. by R. Filler, American Chemical Society Symposium Series No. 28, ACS, Washington, D.C.

13. Ophaug, R. H. and Singer, L. 1980. "Metabolic Handling of Perfluorooctanoic Acid in Rats". Proc. Soc. Exp. Biol. Med., 163:19-23.

14. Rozner, M. A. 1981. "The Effects of Perfluoro-octanoic Acid Upon Microsomal Membrane Peroxidation", Dissertation Abstracts International, 1, No. 08.

15. Griffith, F. D. and Long, J. E. 1980. "Animal Toxicity Studies with Ammonium Perfluorooctanoate", Am. Ind. Hyg. Assoc. J., 41:576-583.

16. Bacon, J., Keller, W., and Andersen, M. 1981. "Teratologic Evaluation of a Model Perfluorinated Acid, NDFDA", Air Force Aerospace Medical Research Laboratory, Wright-Patterson AFB, OH AFAMRL-TR-81-14, (AD A095424).

17. Andersen, M. E., Baskin, G., and Rogers, A. 1981. "The Acute Toxicity of Perfluoro-N-Decanoic Acid: Similarities with 2,3,7,8,-Tetrachlorodibenzodioxin", The Toxicologist, 1:6.

18. Andersen, M., Olson, C., Rogers, A., and Van Rafelghem, M. 1981. "Toxicity of Perfluorinated Fatty Acids in Comparison to that of 2,3,7,8,-Tetrachlorodibenzodioxin", presented at the Review of Air Force Sponsored Basic Research in Environmental Toxicology", The Ohio State University, Columbus, OH 2-3 June.

19. Olson, C. T. 1982. "Effects of Nonadecafluorodecanoic Acid on Tissue Fatty Acids of the Rat", Ph. D. Dissertation, The Ohio State University, Columbus, OH.

20. Fox, J., 1981. "Fatty Acid's Spontaneous Oxidation Clarified", Chem. Eng. News. 18-19, Oct 26.

21. Porter, N. A., Lehman, L. S., Weber, B. A., and Smith, K. J. 1981. "Unified Mechanism for Polyunsaturated Fatty Acid Autoxidation. Completion of Peroxy Radical Hydrogen Atom Abstraction, β-Scission, and Cyclization", J. Am. Chem. Soc., 103:6447-6455.

A PRELIMINARY STUDY OF ELECTROMIGRATION IN ALUMINUM AND ALUMINUM- SILICON FILMS

by

R. L. REMKE & K. T. WILSON

ABSTRACT

Accelerated tests were developed and the experimental apparatus constructed to study the electromigration of aluminum and aluminum-silicon thin films. The accelerated tests were developed to investigate the effect of hydrogen ambient, temperature cycling, and a pulsed current with a direct current bias on electromigration. To perform these tests, a special test chamber was constructed to provide a controlled ambient at elevated temperatures and still allow the electromigration to be monitored by measuring the conductor resistance. Film annealing and step stressing procedures were begun on the thin film aluminum conductors to measure the thermal coefficient of resistance and to estimate the electromigration activation energy. Recommendations for continuing the research are offered.

ACKNOWLEDGEMENTS

The authors would like to gratefully acknowledge the invaluable assistance and gracious hospitality rendered by the Product Evaluation Section of the Reliability Branch at Rome Air Development Center (RADC) in Rome, NY, and in particular the Section Chief, Dr. Robert Thomas. A special thanks to Clyde Lane of the Reliability Physics Section for his guidance and direction in developing the electromigration tests. We would also like to thank David W. Harris for his assistance in constructing the electromigration test chamber, Donald Calabrese for the SEM pictures, and Ivars Irbe for the optical pictures.

Finally, the authors would like to acknowledge the sponsorship of the Air Force Systems Command, Air Force Office of Scientific Research (AFOSR) and the Southeastern Center for Electrical Engineering Education (SCEEE).

I. INTRODUCTION

The electromigration of thin film conductors have been identified as one of the dominant failure mechanisms in integrated circuits.[1-5] Electromigration is the movement of atoms in a metallic conductor driven by the "electron wind" of the current through the conductor. Electromigration can cause voids and hillocks to form in a thin film conductor resulting ultimately in a conductor and circuit failure. The failure rate as measured in the mean time to failure (MTF) has been extensively studied and found to be related to temperature, current density, sample geometry, physical characteristics of the film and substrate, and protective coating.[5] To enhance the reliability of the conductors, new conductor materials have been developed with film properties that are more resistant to electromigration than the widely used aluminum conductors. These materials include aluminum-silicon,[6] aluminum-copper,[7-9] and aluminum-silicon-copper.[6] Work has also continued to determine how the film properties relate to the electromigration mechanism and how the electromigration is affected by ambient, film processing, current density, and thermal stressing.

The present project is concerned with studying the electromigration of aluminum and aluminum-silicon conductors. The effect of a hydrogen ambient, temperature cycling, and a pulsed current with a direct current bias has been chosen for study because of their relevance to the understanding of electromigration and their importance in the operation of integrated circuits. To measure the effect of electromigration, the resistance of the thin film conductor is monitored. This parameter was chosen since it is likely that resistance changes in conductors may destroy or significantly degrade the circuit performance long before the conductor fails. This resistance technique has been previously used to monitor the electromigration of aluminum conductors where the resistance change was related to ion velocity.[10]

A hydrogen ambient has been shown to change the rate of electromigration in aluminum films. Although a detailed understanding of the effect of hydrogen does not exist, studies have shown that hydrogen significantly reduced the electromigration of unpassivated films in accelerated tests and caused a dramatic increase in the mean time to failure (MTF) of the films.[11-12] To more accurately determine the effect of hydrogen, the rate of electromigration as measured by the activation energy and other rate parameters should be determined. The determination of these parameters will provide insight into the effect of hydrogen on the electromigration of thin film conductors and should shed some light on the electromigration mechanism in general.

During the life of an integrated circuit, the thin film conductors may be exposed to temperature cycling. This cycling may be due to variations in ambient temperature or result from the internal Joule heating of the film. The effect of this temperature cycling on electromigration has not previously been studied although certain results suggest that cycling may alter the rate of electromigration in aluminum conductors. Thermal cycling of unpowered aluminum conductors showed changes in the conductor resistance and emissivity.[13] Other work attributed the increased failure rates in aluminum-copper conductors to thermal cycling during pulse conditions.[14] These results imply that thermal cycling may affect the properties of thin film conductors and significantly change the electromigration.

The current through a thin film conductor in a digital integrated circuit can usually be expressed as the combination of a continuous direct current and pulsed current. The effect of a direct current on

electromigration has been thoroughly investigated yielding a MTF with J^{-n} dependence where J is the current density and n varies from 1 to 3. The effect of a pulsed current has also been investigated in electromigration studies. These studies indicated that the pulse duty factor and the increased Joule heating resulting from the pulse significantly affect the electromigration of thin film conductors.[15,16] In this study, we propose to study the effect of the combination of the direct current and pulsed current as described below.

II. OBJECTIVES

The principal objective of this project was to investigate the electromigration of aluminum and aluminum-silicon thin films. The specific project objectives were:

(1) Develop an accelerated test procedure to measure the effect of hydrogen ambient, temperature cycling, and a pulsed current with a direct current bias on electromigration.

(2) Construct a test chamber to perform the electromigration experiments.

(3) Perform and evaluate the electromigration experiments.

During the time at Rome Air Development Center (RADC), the first two objectives were accomplished as described below. Initial work was begun on the third objective with continuation of the electromigration experiments recommended as follow-up work.

III. ACCELERATED ELECTROMIGRATION TEST PROCEDURES AND RESULTS

The accelerated electromigration test procedures are given below. Since an objective of this project was to develop the accelerated test

procedures, the entire procedure is given below although all the electromigration tests were not completed. The uncompleted tests are recommended for follow-up work.

The samples used in this study were 99.999% aluminum or aluminum-silicon (1.5% nominal) thin film conductors provided by RADC. The conductors were 8.2 μm wide and had a thickness of 0.8 ± 0.025 μm. The substrates were silicon-on-sapphire (SOS) where the silicon had been completely steam oxidized at 875°C. To delineate the thin film conductors, the samples were wet etched using a phosphoric, acetic, nitric, water etch at 50°C for 1.5 to 2.0 minutes. The photoresist was then stripped using an acetic, isopropanol, water solution. No passivation was used for these samples.

To perform the electromigration tests in the desired atmosphere, a vacuum test chamber was constructed. This chamber is shown in Figure 1 along with the associated electronics. The chamber consists of a stainless steel cross with appropriate fittings for electrical connections, thermocouple wires, and gas supply. Since the electrical connections must function at 250°C, Teflon coated wire, high temperature solder connections, and appropriate ceramic insulators were used. The chamber is heated with a wrapped heating coil and powered by a Variac.

Optical and scanning electron microscope pictures were taken of the thin film conductors. The optical photographs were taken to identify any variations in the conductor widths before the electromigration tests were run. A typical optical picture is shown in Figure 2. Scanning electron microscope (SEM) photographs were also taken of the conductors to initially view the surface of the films as shown in Figure 3. The aluminum metallizatio was very consistent and smooth while the aluminum-silicon film showed a random distribution of finely distributed particle-like features. These

pictures will be compared with pictures taken during and after the electromigration tests to determine the effect of the electromigration.

Prior to beginning the accelerated tests, the thin film conductors were annealed and the thermal coefficient of resistance (TCR) determined for each conductor. Annealing of the conductors releases the stress incurred during film preparation and allows the TCR to be reliably determined. The annealing and TCR determination were both accomplished by cycling the conductor from room temperature to 250°C until a repeatable resistance vs. temperature curve was obtained. The TCR was used to estimate the actual conductor temperature during the electromigration experiments. TCR values were usually in the range of .03 /°C and linear with temperature.

Step stress studies on the thin film conductors were initiated to provide an estimate of the activation energy and other rate parameters. For these studies, the time (t) required for a specified change in resistance due to electromigration is assumed of the form,

$$t = PE \exp (E/kT) \qquad (1)$$

where PE is the pre-exponential term, E is the activation energy, k is the Boltzmann constant, and T is temperature. It is noted that Eq. 1 can be rewritten as

$$\ln t = E/kT + \ln PE \qquad (2)$$

which yields a straight line on Arrhenius plots. In the step stress procedure, the sample is heated under stress (current) at elevated incremental temperatures. In this study, 150°C, 175°C, 200°C, and 225°C were chosen. The sample is left at each temperature until a predetermined change in resistance is measured. The time is recorded for each temperature and an Arrhenius plot can be constructed to yield the activation energy and pre-exponential term.

The experiments described below were not finished while at RADC but are recommended as follow-on work. These experiments should provide a framework for a comparison of the effect of the parameters previously described on the electromigration of the thin film conductors.

To establish a baseline of comparison for the electromigration tests, the aluminum and aluminum-silicon samples should be tested in an atmosphere of nitrogen. A slightly pressurized non-flowing ambient should be used. The change in resistance should be measured at 175°C, 200°C, and 225°C for a current density of $1X10^6$ A/cm^2. From this data, Arrhenius plots should be used to determine the activation energy and pre-exponential terms previously discussed (See Eqn. 1). Before, during, and after the tests, optical and SEM analysis should be performed on the samples as previously noted.

To characterize the effect of hydrogen on electromigration, the electromigration tests should be repeated using a hydrogen atmosphere. The same test chamber and test procedure should be used as with nitrogen atmosphere. These results should be compared with those obtained from the nitrogen ambient results.

The effect of temperature cycling on electromigration in the aluminum and aluminum-silicon samples should be investigated. Before the electromigration tests begin, the sample should be thermally cycled from -50°C to + 250°C for 300 cycles. The electromigration tests should then be run in a nitrogen atmosphere at 175°C with a current density of $1X10^6$ A/cm^2.

The effect of a pulsed current with a direct current bias should also be studied. Two groups of aluminum and aluminum-silicon samples should be stressed with a direct current density of $3X10^5$ A/cm^2 rms. One group should use a pulse frequency of 1MHz while the other group should use a frequency of

10 MHz. The electromigration tests should be run on both sets of samples in a nitrogen atmosphere at 175°C.

III. RECOMMENDATIONS

During the time at RADC, the electromigration test procedure was developed, the experimental apparatus was set up, and the initial annealing tests were performed as previously described. We recommend that the electromigration study be continued and the effect of the following parameters be determined:

(1) a hydrogen ambient

(2) temperature cycling

(3) a pulsed current with a direct current bias

These parameters can significantly affect the electromigration of thin film conductors and the results of the study should provide insight into the electromigration mechanism.

REFERENCES

1. P. B. Ghate, "Some Observations on the Electromigration in Aluminum Electromigration in Aluminum Films," Appl. Phys. Lett., 11, 14 (1967).

2. I. A. Blech and E. S. Meieran, "Electromigration in Thin Aluminum Films," Appl. Phys. Lett., 11, 263 (1967).

3. J. R. Black, "Electromigration - A Brief Survey and Some Recent Results," IEEE Trans. Electron. Devices, ED-16, 338 (1969).

4. J. J. Bart and M. Walter, "Assurance Procedures for the Control of Electromigration," Proceedings of 20th International Reliability Physics Symposium, 285 (1982).

5. P. B. Ghate, "Electromigration - Induced Failures in VLSI Interconnects," Proceedings of 20th International Reliability Physics Symposium, 292 (198

6. A. J. Learn, "Electromigration Effects in Aluminum Alloy Metallization," J. Electron. Mater., 3, 531 (1974).

7. I. Ames, F. M. d'Hevrle, and R. E. Horstmann, "Reduction of Electromigrat in Aluminum Films by Copper Doping," IBM J. Res. Develop., 461 (1970).

8. F. M. d'Hevrle, N. G. Ainslie, A. Gangvlee, M. C. Shine, "Activation Energy for Electromigration Failure in Aluminum Films Containing Copper," J. Vacuum Sci. and Tech., 9, 289 (1972).

9. I. A. Bleck, "Copper Electromigration in Aluminum," J. Appl. Phys., 48, 473 (1977).

10. R. E. Hummell, R. T. Dehoff, and H. J. Geier, "Activation Energy for Electrotransport in Thin Aluminum Films by Resistance Measurements," J. Phys. Chem. Solids, 37, 73 (1976).

11. D. Shin and P. J. Ficalora, "The Reduction of Au-Al Intermetallic Formation and Electromigration in Hydrogen Environments," Proceedings of 16th International Reliability Physics Symposium, 262 (1978).

12. A. K. Murthi and V. Sheuchin, "The Effect of Hydrogen Ambients on Failure Mechanisms in CMOS Metallization," Proceedings of 20th International Reliability Physics Symposium, 55 (1982).

13. A. C. Macpherson, W. H. Weisenberger, H. M. Day, and A. Christou, "Effects of Fast Temperature Cycling on Aluminum and Gold Metal Systems," Proceedings of 13th International Reliability Physics Symposium, 113 (1975).

14. E. Kinsbron, C. M. Melliar-Smith, A. T. English, and T. Chynoueth, "Failure of Small Thin Film Conductors Due to High Current-Density Pulses," Proceedings of 16th International Reliability Physics Symposium, 248 (1978).

15. A. T. English, K. L. Tai, and P. A. Turner, "Electromigration in Conductor Stripes Under Pulsed DC Powering," 21, 397 (1972).

16. R. J. Miller, "Electromigration Failure Under Pulse Test Conditions," Proceedings of 16th International Reliability Physics Symposium, 241 (1978).

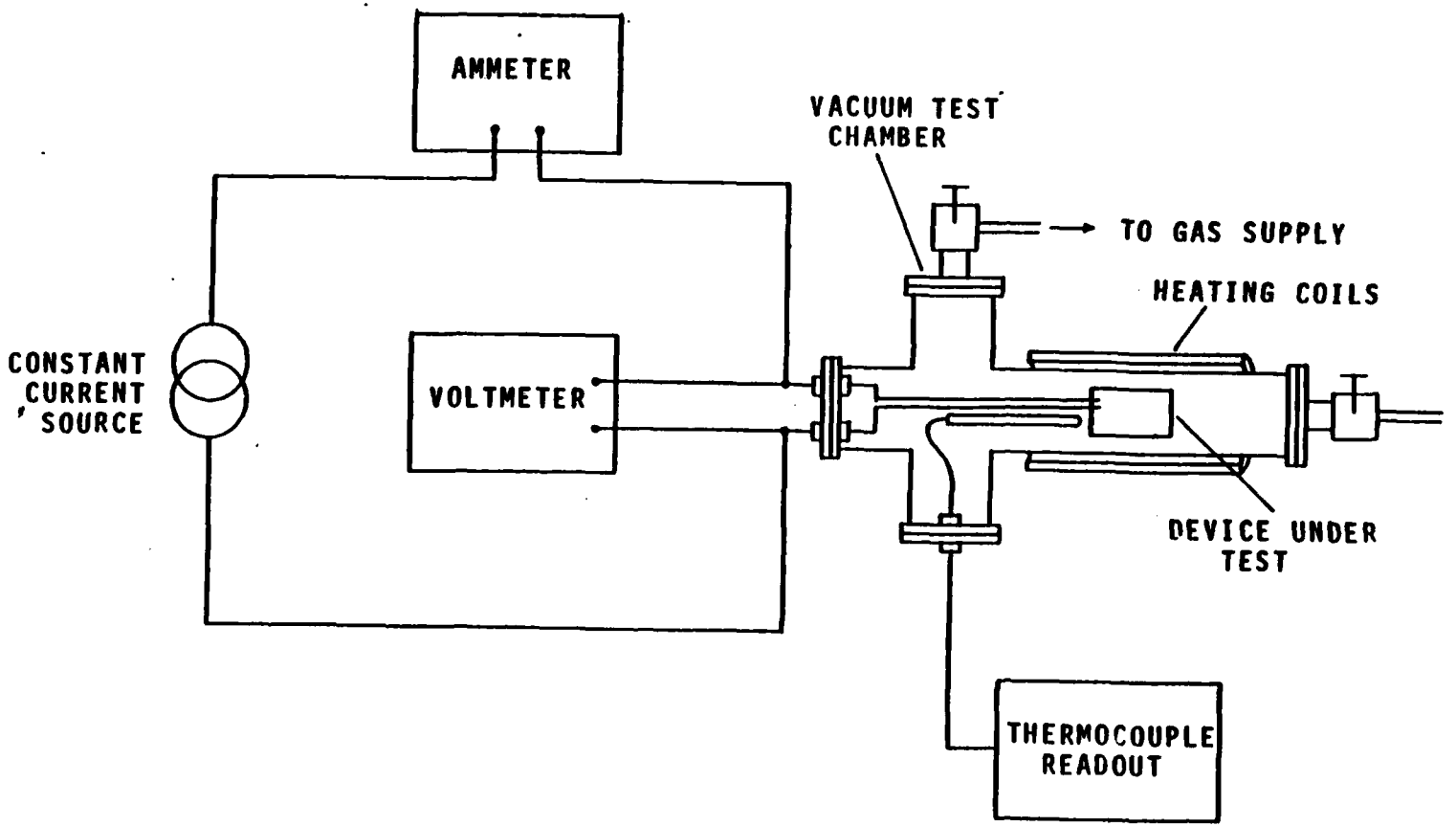


Figure 1. Electromigration test chamber.

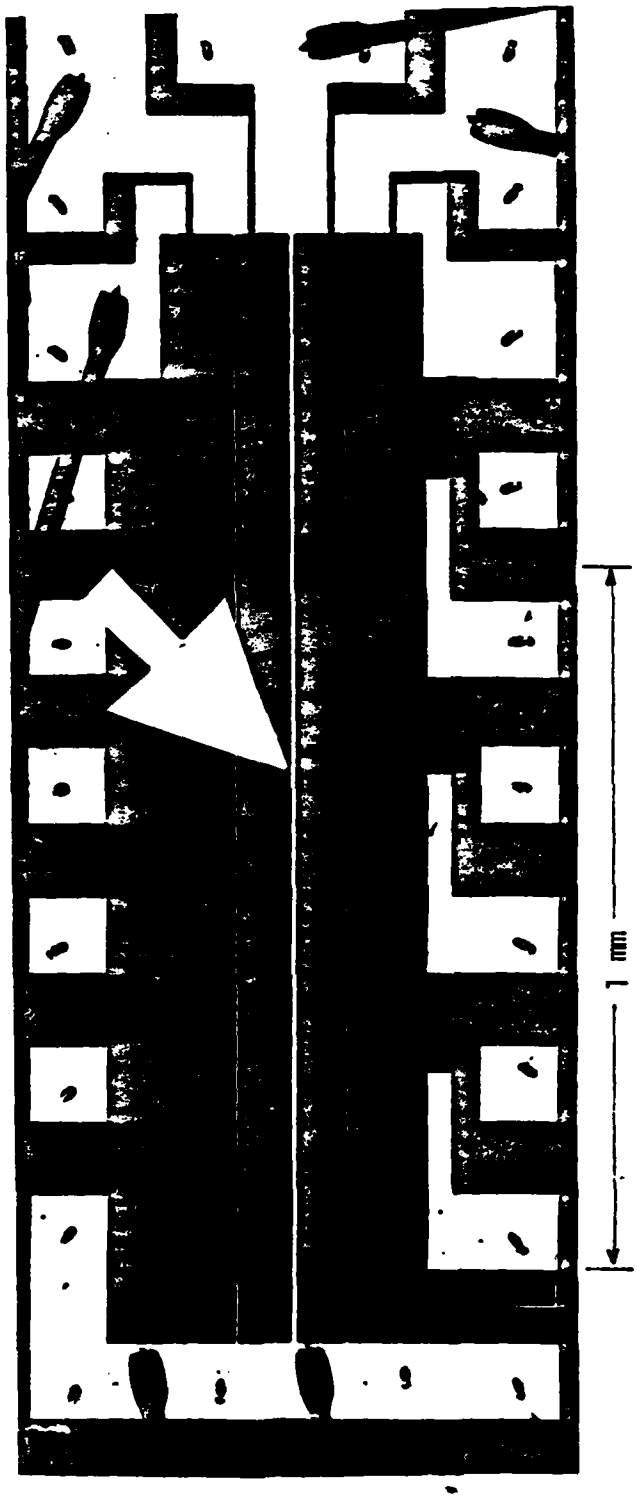


Figure 2. Optical photograph of aluminum thin film conductor.

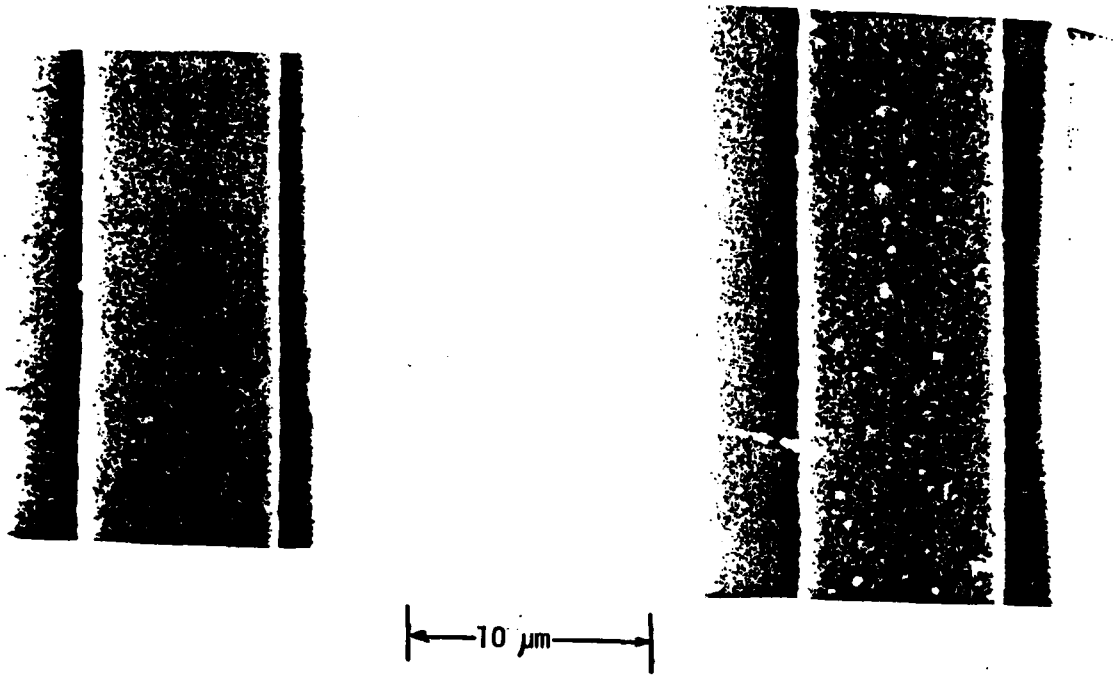


Figure 3. Scanning electron microscope photographs of (a) aluminum (b) aluminum-silicon thin film conductors.

1982 USAF - SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

A STUDY OF FACTORS AFFECTING SOOT FORMATION IN A SWIRL-STABILIZED COMBUSTOR

Prepared by: Dr. Richard W. Rice

Academic Rank: Assistant Professor

Department and

University: Chemical Engineering

Clemson University

Research

Location: Air Force Engineering and Services Center

Environics Division, Environmental

Sciences Branch

USAF Research

Colleague: Major John T. Slankas

DATE: August 9, 1982

Contract No: F49620-82-C-0035

A STUDY OF FACTORS AFFECTING SOOT FORMATION IN A SWIRL - STABILIZED COMBUSTOR

by

Richard W. Rice

ABSTRACT

The effects of a variety of combustion and sampling variables on soot concentration and particle size distribution were studied using a swirl-stabilized laboratory combustor and an Electrical Aerosol Analyzer (EAA). Tests were conducted over a range of reference velocities (2.5 to 10 m/s) and equivalence ratios (0.05 to 0.3) for five gaseous fuels: propane, n-butane, ethylene, propylene, and 1-butene. Soot concentration and size were found to decrease with increases in either gas reference velocity or, surprisingly, equivalence ratio. Alkenes produced more and larger soot particles than alkanes of corresponding carbon number. Sooting tendency increased with increased carbon chain length. Probe placement and sample line length affected soot concentration measurements. Smoke number readings were obtained for comparison with EAA data.

Recommendations were made for further work on the topics lightly surveyed in this study and for extensions which could possibly answer important questions raised here concerning soot growth/oxidation phenomena.

ACKNOWLEDGEMENTS

I would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research, and the Southeastern Center for Electrical Engineering Education for providing the opportunity for me to spend a rewarding summer at the Air Force Engineering and Services Center, Tyndall AFB, Florida. All members of the Environics Division have my gratitude for thier friendly cooperation.

Special thanks go to Major Tom Slankas for his guidance in selecting and conducting this research; to Lieutenant Jeff Jenkins and Captain Dan Berlinrut for their indispensable assistance in actually carrying out the laboratory work; and to Lieutenant Bob Beggs and MSgt Dan Stork for their timely support.

I. INTRODUCTION

The U. S. Air Force has a multi-fold interest in investigating the factors that affect combustion and soot formation in gas turbines. First, from a tactical standpoint, it is desirable to reduce the visibility of the exhaust plume from jet engines. Second, incandescent soot particles have high emissivity and transfer radiant energy from the flame to the jet engine combustor liner. This influences both the liner lifetime and the cooling air requirement, a factor which, in turn, affects engine performance. Finally, there is concern about the environmental impact of jet engine emissions, particularly during static engine tests. This last topic is important not only for aesthetics, but also for health reasons; because most soot particles are of the proper size (0.005 to 0.1) to be ingested deep into the lungs (1). It is also believed that the polynuclear aromatic hydrocarbons comprising or adsorbed on soot (2) are carcinogenic (3).

Several investigations of jet engine emissions have reached the conclusion that combustor design modifications are the most effective method of controlling emissions (4, 5), but the use of soot-suppressing fuel additives has recently received interest (6,7). The chemical properties of aviation fuels have also been shown to have a large influence on sooting tendency. In particular, hydrogen to carbon ratio and weight percent hydrogen have been used to correlate feed properties and soot-related measurements such as flame radiation and opacity (8,9). The polycyclic aromatic content of a fuel has also been cited for its impact on soot formation (10). In addition to fuel properties, combustion conditions, e.g., temperature, fuel to air ratio, velocity, etc., have a major effect on the quantity and nature of soot produced.

Increasing temperature has been shown to reduce soot formation in premixed flames. This is believed to be due to the fact that, as temperature increases, the OH radical concentration in the flame increases more rapidly than that of soot precursors and these hydroxyl radicals are especially effective in the oxidation of soot precursors (11). However, an opposite effect of temperature has been reported in tests utilizing other combustor configurations (12). In diffusion flames fuel pyrolysis is believed to be the dominant process in soot formation, and the ultimate soot concentration in the exhaust is dependent on the retative global reaction rates and characteristic times for the competing pyrolysis and oxidation reactions. When gas reference velocity is increased, flame speed reduces the time available for pyrolysis and increases the probability of oxidizing soot precursors. In laminar diffusion flames, where the time for pyrolysis is long, molecular structure has been found to be more important than H/C ratio (12). For highly turbulent flames, where pyrolysis is rate - limited, the opposite is true, thus it is questionable whether results from laminar flame studies should be used to predict turbine engine behavior. Only modest effects on sooting have been found (10) for fuel to air ratio for lean flames ($F/A \cong 0.01$).

Most studies of soot formation have involved relatively complex molecules or mixtures, e.g., actual jet fuel, and indirect methods of soot characterization e.g., flame emission, opacity, and Smoke Number measurements (10, 13, 14). While valuable, such studies do not provide: 1) information concerning the soot particle size distribution, 2) accurate measures of soot concentration, and 3) information on the sooting behavior of specific compounds. Such data are needed if a

fundamental understanding of the roles played by combustor parameters and molecular structure is to be obtained. The investigation described here hopefully represents a first step in this direction.

II. OBJECTIVES

Objectives for this project existed on three levels. On the "most general" level was the goal of familiarizing this author with (for him) a new research topic, i.e., a sort of combustion/aerosol primer. On the next level there was the general long-range objective of gaining an understanding of the soot formation mechanism in jet turbines by relating laboratory combustor results to operating variables. Finally, the immediate objective was to determine the effect of specific system parameters such as fuel structure, reference velocity, equivalence ratio, sampling system design, etc., on soot concentration and average particle size.

III. EXPERIMENTAL PROCEDURES

The experiments performed this summer involved a variety of equipment. A brief description of each of the major items will be given below, followed by a discussion of how they were used together in a typical case.

Laboratory Combustor

All experiments were made using a swirl-stabilized laboratory combustor scaled down and modified by G.S. Samuelsen and his co-workers at the University of California, Irvine, from an original design used at the Air Force Aero Propulsion Lab. A detailed description of that combustor has been given elsewhere (15), thus only the most pertinent features of the scaled-down version will be described here. The combustor consisted of a 100 cm length of 5.7 cm o.d., 5.2 cm i.d., glass

tubing attached by gasketed flanges to a stainless steel main housing which contained a 3 cm o.d. centerbody that extended 6 cm axisymmetrically away from the flange. Fuel gas was injected into the glass tube through a 0.13 cm i.d. jet in the centerbody. Air was introduced through an annular drilled distributor plate surrounding the centerbody. A swirl was imparted to the air flow by a stainless steel swirl vane encircling the centerbody and equipped with blades slanted at 30°, 45°, or 60° (45° unless otherwise noted). The swirl vane was fitted snugly inside the glass tube 3 cm from the centerbody face. This geometry resulted in rather complex aerodynamics, including a tuburlent annular "sheath" of air surrounding the combustion gases and a strong reverse flow near the radial center of the tube.

Pitot tube measurements taken at the mouth of the glass tube during cold flow tests confirmed the presence of a reverse flow at the center of the tube. This effect was also observed during studies at the University of California, Irvine (16). However, they found that the radial velocity profiles were considerably different during combustion runs. At a reference velocity of 7.5 m/s and an equivalence ratio of 0.05 the centerline velocity near the exit plane was negative during cold flow run, but positive during combustion runs. At a reference velocity of 15 m/s and an equivalence ratio of 0.05 a negative centerline velocity was still observed during combustion. Supporting equipment for the combustor included a rotameter and associated valves for metering gaseous fuel and a calibrated orifice/water manometer system for metering dry compressed air.

Aerosol Analyzer

A Thermo Systems Inc. Model 3030 Electrical Aerosol Analyzer (EAA)

interfaced with an Apple II microcomputer and connected to an Integral Data Systems Model 445 Printer provided measurements for both soot concentration and average particle diameter, as well as bar graphs showing particle size distribution. Three calculational bases were used: 1) number of particles/cm^3, 2) aerosol surface area in μ^2/cm^3, and 3) aerosol volume in μ^3/cm^3. These will hereafter be represented by the symbols N, S, and V respectively. The method of data reduction used by this system is Kapadia's (17) application of a statistical technique originated by Twomey (18). Details of the data reduction and operation of the EAA can be found elsewhere (19). Flow of air and aerosol to the EAA, induced by a vacuum pump, was measured using a mass flow indicator and adjusted to specifications using metering valves. The aerosol sample (4 +/- 0.2 SLPM) passed through a diffusion drier and an aerosol charge neutralizer (Kr-85 source). The EAA was operated in the 9 channel mode in which it detected particles in the 0.01 to 1.0μm size range. This is one of very few means of accurately measuring aerosol size distribution over this particle size range.

General Operating Procedure

Typically, after the room exhaust fan had been turned on, fuel flow was established at the desired value and air flow was initially set at a very low value to facilitate ignition. Next, a small propane torch, was lit and its flame was directed toward the combustor tube mouth. Following ignition, air flow was rapidly raised to the desired value to prevent "burn-back" onto the centerbody. In most instances the sample probe used was a 0.32 cm i.d. stainless steel nozzle (bent into the shape of a hook), which ordinarily would be used with a Hill In-Stack Air

Sampler/Cascade Impactor. The other sample probe used is a water cooled probe with inlets for NO_2 dilution. Unless otherwise stated, the sample line used was a 5 foot length of heated (150°F) 1/4 inch teflon tubing. On several occasions, Smoke Number measurements were made using a Roseco Model 473A Engine Smoke Emission Sampler and a Welch Densichron reflectance meter. The procedure followed was that set forth in the Aerospace Recommended Practice Manual (ARP 1179).

IV RESULTS AND DISCUSSION

Given the relatively short research period available and the large number of potentially important parameters, it was decided that the optimum approach would be to first do a brief determination of the effect of extraneous, e.g., sampling-related, factors; then to study two or three combustion parameters in detail. Specifically, the work was divided into two main areas, each with several subtopics as shown below.

1. Equipment Testing and Sensitivity Studies.
    a. Sample probe type and location
    b. Sample line length and temperature
    c. Sample dilution
    d. Fuel jet (orifice) size
    e. Swirl vane angle
2. Combustion Parameter Studies
    a. Reference velocity, v
    b. Equivalence ratio, ϕ
    c. Chemical nature of the fuel
    d. Fuel additives

Sample Probe Type and Location

As mentioned earlier, crude pitot tube measurements of gas velocity

profile showed a strong reverse flow near the tube center and a strong forward flow near the tube wall. Numerous tests were made to determine the extent to which this flow pattern affected soot measurements. For a relatively low velocity case, Figure 1 shows that at the tube mouth (axial distance = 0 inches) the aerosol concentration was only about 1/3 that near the wall, while only 2 inches outside the mouth of the tube (axial distance =-2) the concentration, N, was nearly independent of radial position. Table 1 contains additional data which show that probe location mainly affected concentration, with S and V being more sensitive than N. Note that average particle size was only slightly dependent on probe position, with somewhat larger particles prevalent near the wall. These data also illustrate why the probe was generally located at the radial center and 3.5 inches away from the tube, i.e., (-3.5, 0), because at that axial location, EAA results were not sensitive to radial location.

Table 2 shows that, although there may have been some differences in EAA measurements when using a 0.63 cm i.d. impactor probe rather than the standard 0.32 cm i.d. probe, these were relatively small and no consistent pattern was observed. It should be noted that the 0.32 cm probe was chosen for general use because calculations indicated that it should most closely approximate isokinetic sampling for the gas velocity range used.

Sample Line Length and Temperature

Table 2 also illustrates that tripling the length of the sample line essentially halved the concentration of the sample reaching the EAA. Losses due to impaction on the sample line walls were modest, as might be

expected for such small particles. There was a small shift in size distribution to lower diameter as line length increased. This presumably reflects the higher probability that large particles would be captured by the walls. In making these measurements it was necessary to adjust the EAA Aerosol Flow metering valve to offset the increased flow resistance of the longer sample lines. Tests showed no appreciable effect of sample line temperature until above 200°F, where the EAA Analyzer Current was typically 45% higher at 250°F than at 150°F. However, significant transients were observed when sample line temperature was increased, possibly indicating a desorption effect.

Sample Dilution

Once the aerosol concentration exceeded roughly 10^6 particles/cm^3, the EAA would generate an error message indicating failure to obtain a satisfactory iterative fit to the data. This condition was encountered for almost all samples from combustion of alkenes. In such cases it was necessary to dilute the "true" aerosol sample by introducing a known flow of N_2 through a tubing tee located near the probe. Several tests were made to determine: 1) which of the three concentration measures N, S, or V, was best correlated with % exhaust gas in the sample (=100% - % dilution) and 2) whether dilution caused a major shift in size distributuion due to condensation or other effects.

First, because of the sensitivity of EAA data to the aerosol gas sample flow rate (nominally 4 SLPM) and the day to day shift in this quantity, it was found necessary to periodically determine the dilution gas flow rate corresponding to virtually complete dilution. To so this, EAA Analyzer Current was measured at increasing dilution gas flow rates

until it approached the "background" value, typically 0.01 to 0.04. Then a curve was fit to the data and extrapolated to this nearly zero value. Figure 2 shows two such plots, with values of 3.7 and 4.5 SLPM for 100% dilution. Figure 3 shows the variation of normalized concentration (diluted gas value/undiluted gas value) on all three bases versus % exhaust gas in the sample for n-butane at ϕ=0.1 and v =7.5 m/s. A nearly identical plot was also obtained for propane at ϕ=0.3, v =3.5 m/s, so Figure 3 is representative of dilution behavior over a wide range of variables. The most obvious feature of this figure is that S appears to vary in an almost perfect linear manner with % dilution. The other two quantities, N and V, show definite non-linear behavior, thus S is the best basis for use in comparisons where extrapolation of data from one dilution level to another is necessary. A firm explanation for this effect of dilution would require more investigation and it may be a function of dilutor geometry, dilution gas temperature, etc. However, one possible reason why S was the most linear in its behavior may be that the size distributiuon curve, dS/d log D versus D, was often nearly Gaussian, whereas those for N and V were generally more skewed, and thus more affected by dilution - induced changes at the extremities of the distribution. Some of these features can be seen in Figure 5, which shows bar graphs for two cases of dilution representing nearly a factor of 3 difference in % exhaust gas. Figure 4 also illustrates that S varies linearly with dilution for a variety of cases. The dashed line on this plot illustrates that, at high % dilution, it is important to get an accurate value for the hypothetical 100% dilution flow rate mentioned earlier. Ideally, important comparisons of sooting behavior should be made at identical dilution conditions.

Miscellaneous Equipment Studies

A small number of runs were directed at qualitatively checking the effect of fuel orifice size and swirl angle. No appreciable effect on soot characteristics was found for propane when a 0.07cm i.d. fuel jet was substituted for the standard 0.13 cm one. However, the range of ϕ and ν for stable flame operation differed. Similarly, the major effect of using a 60° swirl rather than the usual 45° one appeared to be a more restrictive operating range.

Combustion Parameter Effects

Velocity - Table 3 shows data for two cases which illustrate the effect of nominal (cold flow reference) gas velocity on both soot concentration and size. Case (a) for propylene at ϕ=0.1 shows that increasing velocity over the range, 5.5 to 10 m/s, caused an apparent increase in N, a small decrease in S and a somewhat larger decline in V. For all three calculational bases, the average particle size definitely decreased as velocity increased. The geometric standard deviation of the number distribution, σ_G, also decreased as velocity increased, indicating a narrower size distribution. In case (b) for propane at ϕ=0.3, all measures of concentration decreased significantly as velocity was increased over the range 2.5 to 3.5 m/s. The effect here was much more pronounced than in Case (a), possibly indicating that fuel-rich flames, e.g., Case (b), are more sensitive to velocity variations. Again, there was a significant decrease in average particle size as velocity increased. There may be several potential explanations for these results, but almost certainly one contributing factor is the shorter residence time for all regions of the combustor flame, i.e., pyrolysis,

carbon/hydrogen oxidation, etc., at higher velocity. A shorter time for pyrolysis would result in "nucleation" of fewer soot particles and less growth of any particles formed. Although probably not as important as the pyrolysis time effect, higher velocity would also reduce the time available for particle growth downstream of the visible flame. This could be checked by using combustion tubes of different lengths at a fixed gas velocity. Finally, at higher velocity, there is increased turbulence, and therefore better mixing of fuel, soot precursors, and oxidizer species (such as O_2, $O\cdot$, $OH\cdot$), thus one might expect lower particle counts and sizes due to improved mass transport resulting in an increase in the global reaction rate for oxidation.

Equivalence Ratio - Table 4 presents data for two cases similar to those used in Table 3, but here velocity was held constant and equivalence ratio, ϕ, (defined as actual fuel to air ratio divided by the stoichiometric fuel to air ratio) was varied. For Case (I) as ϕ was increased from 0.05 to 0.15 for propylene, N decreased and d_N increased. Both S and V, as well as d_S and d_V, appeared to reach maximum values around ϕ =0.1. Case (II) shows the behavior most often observed when ϕ was increased, namely, all measures of concentration decreased. In addition, all measures of particle size also decreased, in this instance, substantially. There are at least two factors that are affected when ϕ is increased. First and most obvious, is the fact that the fuel to oxidizer ratio increases, which (everything else equal) might be expected to increase soot formation because more hydrocarbon per unit volume is introduced into the pyrolysis zone and soot precursors would be more likely to encounter and grow by capture of fragmented hydrocarbon species

and less likely to encounter oxidizer species. Indeed, as ϕ was increased, the physical appearance of the flame changed, with the flame becoming longer and brighter yellow, and indication of increased incandescence which would normally be assumed to mean increased soot production. However, this trend was not generally observed, thus one is forced to consider that the dominant effect of increasing ϕ, as it was actually done experimentally, was to increase the fuel velocity at fixed air velocity. This, as just discussed in explanation of the effect of velocity, would decrease the residence time of the fuel in the pyrolysis zone and thus decrease both soot mucleation and growth. However, this explanation is not supported by the small number of results obtained when using the smaller, 0.07 cm, fuel orifice. No major differences in soot characteristics were observed even though the velocity of the fuel as it emerged from the jet would have differed substantially. Of course, this is an overly simplistic picture of the complex fluid mechanics that undoubtedly occur just beyond the centerbody and within the visible flame. In effect, there is no convincing explanation at present for the behavior observed and further experiments are needed.

Chemical Nature of the Fuel - Of the five fuels studied two were alkanes (propane and n - butane) and three were alkenes (ethylene, propylene and 1-butene). These choices enabled two principal comparisons: 1) the effect of degree of unsaturation (related to H/C ratio)0 at fixed carbon chain length, e.g., propane versus propylene; and 2) the effect of chain length for a given molecular type or homologous series. Table 5 presents the results of this study summarized in the form of a "relative sooting tendency" based on n-butane as the reference.

As an example, the relative sooting factor of 4.5 for 1-butene based on N means that, at the specified conditions, 1-butene produced soot with 4.5 times the particle concentration produced by n-butane. Several observations can be made. First, for a given molecular type, sooting tendency increased with increasing carbon chain length, with the effect on V being the most dramatic. Considering the soot-making mechanism involves, among other things, formation of polynuclear aromatics, this trend is understandable since the formation of multi-ring compounds requires the combination of carbon chains. The probability of ring closure is greater for longer chains because these have more potential contact points (carbon atoms) for incorporation into a growing network. A second observation is that, for a given chain length, alkenes produce more soot than alkanes. This is consistent with the picture of soot precursor formation occurring via successive elimination of hydrogen to form highly unsaturated polymeric species. Alkenes, being originally less saturated, might be thought of as already being "partially pyrolyzed". More importantly, the double bond of the alkenes make them more reactive for pyrolysis, ring formation, and/or attachment as an adduct to a growing carbonaceous residue. Interestingly, there is a much larger difference between 1-butene and propylene than between propylene and ethylene, even though the chain length ratio is less (4/3 versus 3/2). In any case, the effect of molecular type is much greater than that of chain length. Given the large differences in combustor conditions for the two cases shown in Table 5, the agreement of relative sooting tendency values is surprisingly good.

Table 6 compares particle size data for the same cases presented in Table 5. Consistent with the previous discussion, an alkene produces

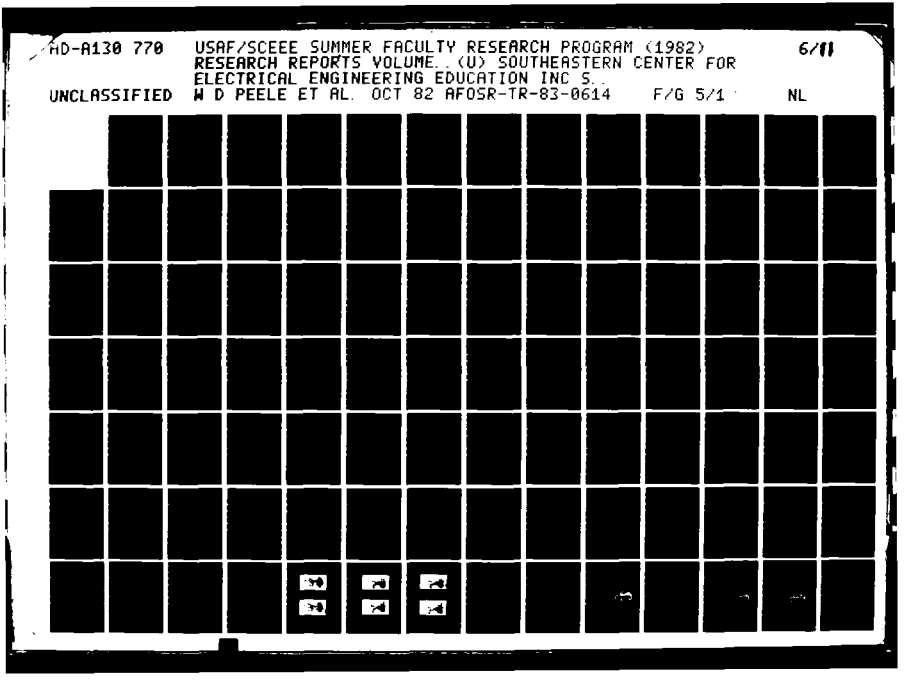
much larger soot particles than its corresponding alkane. The effect of chain length is less obvious here than in Table 5, with 1-butene seemingly producing a slightly smaller average particle than propylene. In this particular case, the differences among these two alkenes are nearly within expected experimental error, and more testing would be needed to obtain statistically significant comparisons. Note that in making such comparisons, particularly where dilution is involved, it is vital that care be taken to insure that all of the multitude of sampling parameters, including EAA aerosol flow and probe location, be held constant.

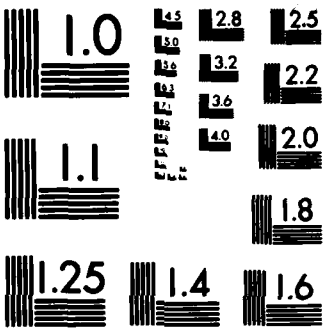
Comparison of EAA and Smoke Number Results

Table 7 presents EAA results (S and d_s) and Smoke Number values for several cases involving propane, ethylene and propylene. The choice of S for the EAA data was based on the fact that as a measure of surface area, it was most closely related to the factors affecting the light reflectance of a soot deposit on filter paper, since that is also surface-related. Furthermore, since the EAA results for the alkenes involved diluted samples while those for Smoke Number did not, S was the obvious choice. As demonstrated earlier, only S can be safely extrapolated to the 0% dilution basis needed here. Given the narrow range of the results, all that can be said is that there is a rough correlation with a given unit of Smoke Number corresponding to an S increment of 600-900 μ^2/cm^3 for the alkenes, but only 350-400 μ^2/cm^3 for propane.

Additional Comments

Two other items deserve mention. One is that gas temperature measurements were occasionally made in the sampling region near the end


AD-A130 770 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM (1982) RESEARCH REPORTS VOLUME..(U) SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION INC S..
6/11
UNCLASSIFIED W D PEELE ET AL. OCT 82 AFOSR-TR-83-0614 F/G 5/1
NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

of the combustor tube using a chromel-alumel thermocouple and a temperature gauge. Some of these results are presented in Table 8 to illustrate: 1) that steep axial and radial temperature gradients existed and 2) that equivalence ratio had a dramatic effect on exhaust gas temperature. Given the large differences in gas temperature near the tube exit, there was less dependence of soot characteristics on probe placement than might have been expected if particle growth by condensation was an important factor in this temperature range. Actually, it is possible that any such effects could have been masked by the effect of the 150°F sample line. The use of a physically non-intrusive measurement technique such as laser-optical scattering might reveal differences that were obscured in this study.

Finally, an attempt was made to obtain soot concentration measurements using a cascade impactor for comparison with EAA results, but this proved unsuccessful. The glass fiber filter elements in the impactor adhered too strongly to the metal support plates, making neat removal impractical. This made accurate determination of the milligram quantities of deposited soot impossible. This problem could probably have been solved but there was not enough time available.

V. CONCLUSIONS

Based on the results of this study the following conclusions were reached.

1. Sample probe location and sample line length have a definite effect on EAA measurement of soot concentration, but a weaker effect on particle size distribution.

2. For a given gas and equivalence ratio, increasing the reference velocity decreases both soot concentration and average particle size,

probably due to decreased residence time for soot nucleation and growth in the pyrolysis zone.

3. Soot concentration and particle size appeared to decrease with increasing equivalence ratio at constant air velocity. This observation runs counter to expectations based on a simple-minded consideration of the probable effect of increased fuel concentration on soot formation. It may be related to the effect of fuel velocity cited above, but no convincing explanation is available, at this time. More experimentation is needed on this topic to "de-couple" the fuel velocity and equivalence ratio effects.

4. Sooting tendency and average particle size increase with the degree of unsaturation of the fuel molecule and, to a lesser extent, with carbon chain length. These trends are consistent with the mechanism of soot formation via growth of a hydrogen-deficient polymer.

5. EAA measurements of soot concentration on a surface area basis, i.e., S_a correlate linearly with % dilution, while N and V do not. For this reason, results for diluted samples are best compared using S.

VI RECOMMENDATIONS

The following are recommendations for future work.

1. The current study was a rapid survey of a multi-variable research area, thus there are many topics only lightly touched here which need further study. Additional runs could determine the importance of extraneous parameters such as: 1) probe type (isokinetic, water-cooled versus simple impactor), 2) combustor tube length, 3) swirl angle and placement on the centerbody, and 4) fuel jet size. The effect of tube length might be two-fold. First it affects the time available for

condensation downstream of the visible flame. Second, due to the strong recirculation caused by the swirl, the proximity of the tube mouth to the flame might be important.

2. The effect of equivalence ratio on soot characteristics should be further studied in order to remove some existing ambiguities. Comparison with similar results using premixed flames could be useful.

3. To continue the work begun on relating sooting tendency to molecular structure, tests should be extended to include such gaseous compounds as isobutane, butadiene, and 1-butyne (ethyl acetylene). The iso-compounds would help separate the effects of carbon number and continuous chain length. Tests of unsaturates such as dienes and alkynes could provide valuable insight into the soot growth phenomenon. Both are prime soot precursors.

4. Tests should be extended to liquid fuels. Soot from single compounds, mixtures of 3 or 4 components, and actual aviation fuel might be compared using the EAA to extend the study of molecular structure on sooting.

5. The effect of various concentrations of additives such as ferrocene, metal naphthenates, and solutions of metal salts such as $Ba(NO_3)_2$ and $BaCl_2$, should be studied by conducting sooting experiments with these aspirated (atomized) into gaseous fuel or dissolved in a liquid fuel.

6. The Laser - Doppler Velocimeter/Particle Sizing System soon to be received at Tyndall AFB should be used to unobtrusively observe the critical soot-making regions in and around the flame. This should better elucidate the detailed interaction between convective fluid mechanics,

molecular transport, and reaction kinetics. It would be particularly interesting to try to measure spectroscopically the relative concentration of important highly reactive (very low concentration) species such as OH radicals and other chemical intermediates.

7. Finally, characterization by chemical means of soot deposited under extremely different conditions might be attempted. Even something as simple as determining the extent to which soot dissolves in a series of solvents could be revealing.

REFERENCES

1. Goldsmith, J. R., in Air Pollution, A. C. Stern, ed., Vol. 1, p. 355, Academic Press, New York (1962).

2. Stoker, H. S., Seager, S. L., and Capener, R. L., Energy, p. 163, Scott Foresman, Glenview, Illinois (1975).

3. Dipple, A., in Chemical Carcinogens, C. E. Searle, ed., p. 245, Amer. Chem. Soc. Monograph 173, Washington, D.C. (1976).

4. Fiorello, S. C., "The Navy's Smoke Abatement Program", SAE Paper No. 680345, April, 1968.

5. Shirmer, R. M., "Effect of Fuel Composition on Particulate Emissions from Gas Turbine Engines", in Emissions from Continuous Combustion Systems, p. 189, W. Cornelius and W. G. Agnew, ed., Plenum Press, N. Y. (1972).

6. Howard, J. B., and Kausch, W. J., Jr., "Soot Control by Fuel Additives, A Review", Air Force Engineering and Services Center Report No. ESL-TR-79-32, September, 1979.

7. Krause, H. H., Hillenbrand, L. J., Wellar, A. E., and Locklin, D. W., Battelle-Columbus report to EPA, EPA Report No. EPA-600/2-77-007a, NTIS No. PB 264068 (Jan., 1977).

8. Naegeli, D. W., and Moses, C. A., "Effect of Fuel Molecular Structure on Soot Formation in Gas Turbine Engines", ASME Paper 80-GT-62 (1980).

9. Vogel, R. E., Troth, D. L., and Verdow, A. J., "Fuel Character Effect on Current High Pressure Ratio Can Type Turbine Combustion Systems", Air Force Aero Propulsion Laboratory Report No. AFAPL-TR-79-2072, April, 1980.

10. Naegeli, D. W., Dodge, L. G., and Moses, C. A., "The Sooting Tendency of Fuels Containing Polycyclic Aromatics in a Research Combustor", AIAA Paper 82-0299, January, 1982.

11. Millikan, R. C., "Non-equilibrium Soot Formation in Premixed Flames", J. Phys. Chem., 66, 794 (1962).

12. Glassman, I., and Yaccarino, P., "The Temperature Effect in Sooting Diffusion Flames", Eighteenth International Symposium on Combustion, The Combustion Institute, p. 1175 (1981).

13. Schirmer, R. M. and Twigg, H. T., "High Pressure Combustor Studies of Flame Radiation as Related to Hydrocarbon Structure", Phillips Petroleum Co. Report No. 3952-65R.

14. Lefebvre, A. H., "Progress and Problems in Gas Turbine Combustion", Tenth International Symposium on Combustion, Univ. of Cambridge, England (1964).

15. Roquemore, W. M., Bradley, R. P., Sturtrud, J. S., Reeves, C. M., and Krishnamurthy, L., "Preliminary Evaluation of a Combustor for Use in Modeling and Diagnostics Development", ASME Paper 80-GT-93, (1980).

16. Brum, R. D., Ikioka, L. M., and Samuelsen, G. S., "Assessment of Candidate Combustor Configurations as Test Beds for Modeling Complex Flows," ASME Paper 82-HT-36 (1982).

17. Kapadia, A., PhD, Thesis, Mechanical Engineering Department, University of Minnesota, Minneapolis, Mn., 1979.

18. Twomey, S., "Comparison of Contrained Linear Inversion and an Iterative Non-Linear Algorithm Applied to the Indirect Estimation of Particle Size Distributions", Journal of Computational Physics, 13, 188 (1975).

19. Liu, B. Y. H., Pui, D. Y. H., and Kapadia, A., "Electrical Aerosol Analyzer: History, Principle and Data Reduction", in Aerosol Measurement, D. A. Lundgren et al, editors, University Presses of Florida, Gainesville, Fla., 1979.

TABLE 1. EFFECT OF PROBE LOCATION ON PARTICLE CONCENTRATION AND AVERAGE PARTICLE DIAMETER

CONDITIONS AND PROBE LOCATION	PARTICLE CONCENTRATION $NX10^{-3}$	S	V	AVERAGE DIAMETER, um $d_N X10^2$	$d_s X10^2$	$d_V X10^2$
PROPANE, ϕ=0.3, V=3.5m/s:	Particles/cm^3	μ^2/cm^3	μ^3/cm^3			
(0,0)	08	174	0.8	2.0	2.6	2.9
(0,-0.4)	158	295	1.5	2.1	2.7	3.2
(0,-0.8)	280	737	4.4	2.5	3.3	3.9
(-2,0)	220	450	2.3	2.2	2.9	3.3
(-2,-0.4)	218	464	2.4	2.3	3.0	3.4
(-2,-0.8)	216	438	2.2	2.2	2.9	3.3
ETHYLENE, ϕ=0.15 V=.5 m/s:						
(0,0) (1)	82	976	21	3.3	10.2	16.2
(0,-1) (1)	349	6599	146	5.3	11.0	16.4
(-3.5,0) (2)	348	4569	90	4.0	9.7	14.5
(-3.5,-1) (2)	433	6481	131	4.5	10.0	15.0

(1) at 85% dilution

(2) at 75% dilution

TABLE 2. EFFECT OF PROBE DIAMETER AND SAMPLE LINE LENGTH ON SOOT CHARACTERISTICS

SAMPLE LINE LENGTH, FEET	$N \times 10^{-3}$ Particle/ cm^3	S μ^2/cm^3	V μ^3/cm^3	$d_N \times 10^2$ um	$d_s \times 10^2$ um	$d_v \times 10^2$ um
5 (1)	516	2090	17.6	3.3	4.6	5.6
10 (1)	390	1709	14.0	3.2	4.5	5.4
15 (1)	282	1001	7.4	2.8	4.1	4.9
PROBE INSIDE DIAMETER, INCHES						
1/8 (1)	426	1240	8.1	2.6	3.6	4.3
1/4 (1)	497	1136	6.5	2.3	3.2	3.8
1/8 (2)	475	3090	34	3.6	5.8	7.3
1/4 (2)	572	4271	47	4.0	6.0	7.4

(1) Propane at ϕ=0.3, V=3.5m/s, Probe at (-2,0)

(2) Propane at ϕ=0.3, V=2.5m/s, Probe at (-2,0)

Table 3: Effect of Gas Velocity on Soot Characteristics

a. Propylene, ϕ =0.1, Probe at (−3.5,0), 73% dilution

Velocity m/s	Analyzer Current	N/1000	S	V
5.5	4.08	261	5129	158
7.5	4.55	348	4998	140
10.0	5.10	461	4768	115

Velocity m/s	$100d_N$ μm	$100d_S$ μm	$100d_V$ μm	σ_G
5.5	3.5	14.7	23.2	2.53
7.5	3.2	12.9	22.1	2.31
10.0	3.0	10.8	19.6	2.11

b. Propane, ϕ =0.3, Probe at (−2,0)

Velocity m/s	Analyzer Current	N/1000	S	V
2.5	6.23	427	5141	76
3.0	5.62	434	4214	55
3.5	3.28	363	1840	17

Velocity m/s	$100d_N$ μm	$100d_S$ μm	$100d_V$ μm	σ_G
2.5	4.9	7.8	10.0	1.62
3.0	4.5	6.9	8.6	1.58
3.5	3.3	5.0	6.1	1.64

TABLE 4. EFFECT OF EQUIVALENCE RATIO ON SOOT CHARACTERISTICS

I. Propylene, V=7.5m/s, Probe at (-3.5,0) 73% dilution

EQUIVALENCE RATIO, ϕ	ANALYZER CURRENT	$N_{x10^{-3}}$ Particles/cm^3	S μ^2/cm^3	V μ^3/cm^3
0.05	4.06	397	3953	102
0.10	4.55	348	4998	140
0.15	3.78	229	4076	106

EQUIVALENCE RATIO, ϕ	$d_n x10^2$ μm	$d_s x10^2$ μm	$d_v x10^2$ μm	σ_G
0.05	2.7	11.6	20.5	2.16
0.10	3.2	12.9	22.1	2.31
0.15	4.7	12.1	20.5	1.97

II. Propane, V=3.5m/s, Probe at (-2,0)

EQUIVALENCE RATIO, Φ	ANALYZER CURRENT	$N_{x10^{-3}}$ Particles/cm^3	S μ^2/cm^3	V μ^3/cm^3
0.11	6.55	433	5433	80
0.20	4.27	364	2916	34
0.30	2.68	309	1473	13

EQUIVALENCE RATIO, Φ	$d_n x10^2$ μm	$d_s x10^2$ μm	$d_v x10^2$ μm	σ_G
0.11	5.0	8.0	9.9	1.65
0.20	4.0	6.3	7.7	1.63
0.30	3.1	4.8	5.9	1.59

TABLE 5. COMPARISON OF RELATIVE SOOTING TENDENCY FOR VARIOUS GASES BASED ON A NORMALIZED VALUE OF 1.0 FOR N-BUTANE*

I. ϕ=0.1, V=7.5m/s, Probe at (-3.5,0)

	RELATIVE SOOTING FACTOR BASED ON			
GAS	ANALYZER CURRENT	NUMBER CONCENTRATION	SURFACE AREA	VOLUME
PROPANE	0.8	0.9	0.7	0.7
N-BUTANE	1.0	1.0	1.0	1.0
ETHYLENE	7	3.0	18	70
PROPYLENE	6	2.5	19	100
1-BUTENE	11.5	4.5	33	150

*-Propane/butane comparison at 0% dilution.

*-Ethylene/propylene/butene/butane comparisons at 90% dilution. All data except for propylene taken on same day.

II. ϕ=0.3, V=3.5m/s, Probe at (-3.5,0)

	RELATIVE SOOTING FACTOR BASED ON			
GAS	ANALYZER CURRENT	NUMBER CONCENTRATION	SURFACE AREA	VOLUME
N-BUTANE	1.0	1.0	1.0	1.0
PROPYLENE	5.3	1.3	21	80
1-BUTENE	12.3	4.7	43	160

TABLE 6. COMPARISON OF AVERAGE PARTICLE DIAMETER FOR SEVERAL FUEL GASES AT COMMON CONDITIONS

I. ϕ =0.1, V=7.5m/s, Probe at (-3.5,0)

GAS	AVERAGE PARTICLE DIAMETER $d_N \times 10^2$ μm	$d_s \times 10^2$ μm	$d_v \times 10^2$ μm	σ_G
PROPANE (1)	2.1	3.2	6.8	1.45
N-BUTANE (1)	2.4	3.6	6.7	1.47
N-BUTANE (2)	1.8	2.7	4.5	1.42
ETHYLENE (2)	3.0	9.0	16.3	2.02
PROPYLENE (2)	3.1	12.2	21.4	2.17
1-BUTENE (2)	3.3	10.7	19.1	1.96

(1) at 0% dilution

(2) at 90% dilution

II. ϕ =0.3, V=3.5m/s, Probe at (-3.5, 0)

GAS	AVERAGE PARTICLE DIAMETER (3) $d_N \times 10^2$ μm	$d_s \times 10^2$ μm	$d_v \times 10^2$ μm	σ_G
N-BUTANE	1.8	2.6	6.0	1.40
PROPYLENE	4.4	14.1	21.8	2.38
1-BUTENE	2.7	13.2	21.9	2.28

(3) at 75% dilution

TABLE 7. COMPARISON OF SMOKE NUMBER AND EAA RESULTS

(Smoke Number Samples were undiluted, EAA samples were diluted as indicated)

RUN NO.	CONDITIONS	SMOKE NUMBER	EAA DATA: ACTUAL S MEASURED μ^2/cm^3	EAA DATA: S, EXTRAPOLATED TO 0% DILUTION μ^2/cm^3	$d_s \times 10^2$ μ
1	Propane, ϕ=0.11, V=3.5m/s, 60° Swirl, Probe at (-2,0), 15 ft. Sample Line	14	5119	5119	8.1
2	Ethylene, ϕ=0.15, V=7.5m/s, Probe at (0,0), 80% dilution	15	---------NOT MEASURED---------		
3	Same as Run 2, but Probe at (-3.5,0)	29	3542	17710	9.5
4	Same as Run 3, but V=5.5m/s	28	4569	22845	9.7
5	Propylene at ϕ=0.1, V=7.5m/s, Probe at (-3.5,0), 5 ft. Sample Line, 78% dilution	28	5389	24495	12.2
6	Same as Run 5, but ϕ=0.3, V=3.5m/s, 73% dilution	28	6201	24804	14.2
7	Same as run 6, but with 15 ft. Sample Line	30	---------NOT MEASURED---------		

TABLE 8. THE VARIATION OF GAS TEMPERATURE WITH PROBE LOCATION AND OTHER COMBUSTOR VARIABLES

CONDITIONS	PROBE LOCATION (inches, inches)	APPROXIMATE GAS TEMPERATURE, °F
Propane at ϕ=0.3, V-3.5m/s	(0,0)	1250
	(0,-1)	600
	(-1,-1)	400
	(-1,0)	850
Propylene at ϕ=0.3 V=3.5m/s	(0,0)	1250
	(-3.5,0)	500
Propylene at ϕ=0.1 V=7.5m/s	(0,0)	550
	(-3.5,0)	150
Propylene at ϕ=0.15 V=7.5m/s	(-3.5,0)	275

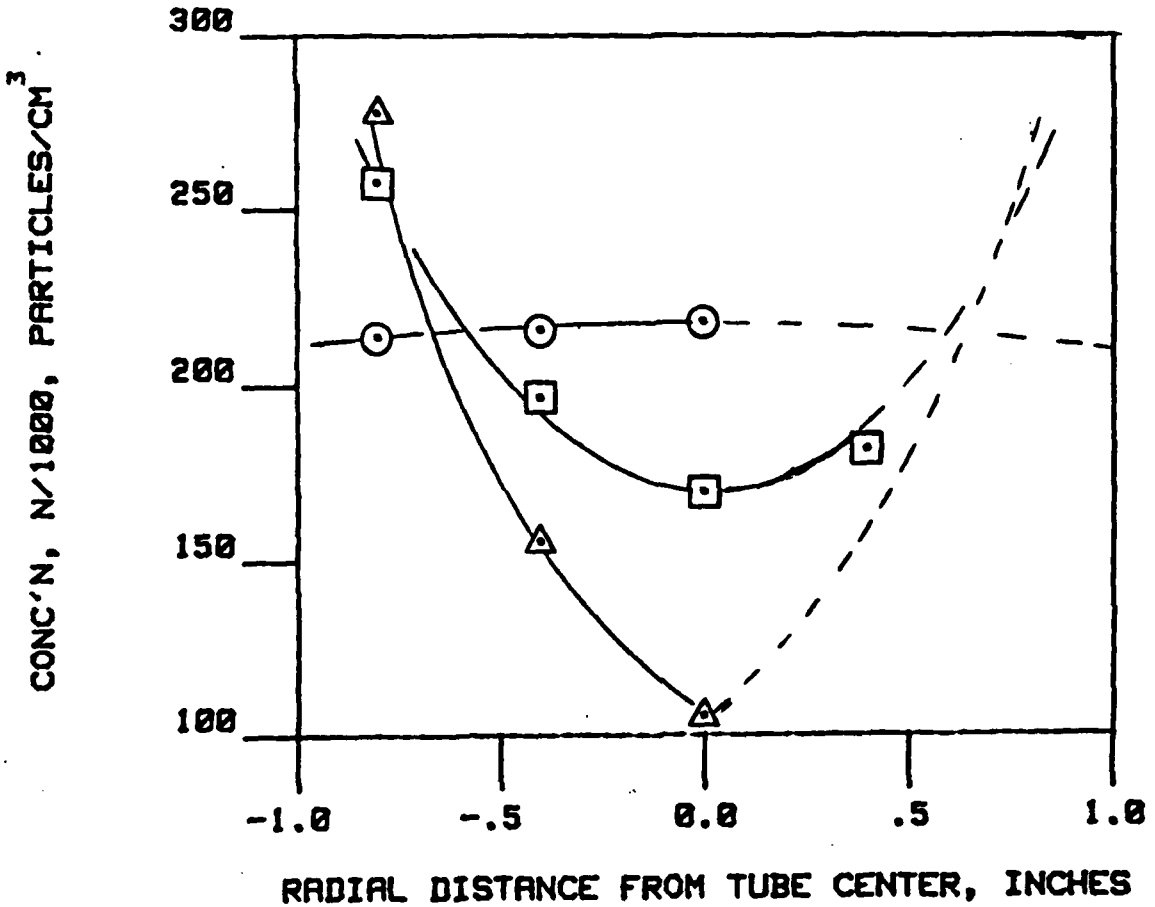


Figure 1. The Variation of Soot Particle Concentration with Probe Location. Propane at ϕ = 0.3, V = 3.5 m/s; △, ⊡, and ⊙ correspond to axial distances of 0, 1, and 2 inches, respectively, away from tube end.

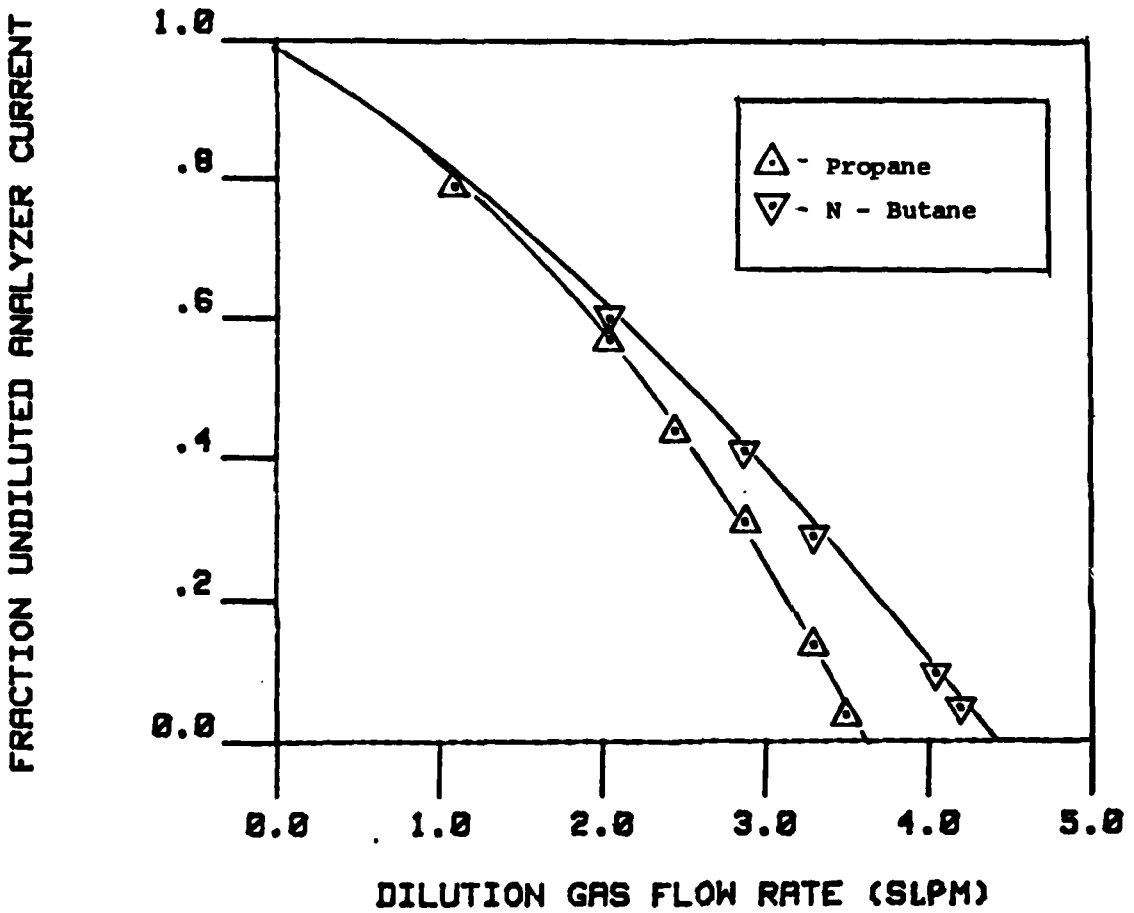


Figure 2. Aerosol Dilution Study; Determination of Hypothetical 100% Dilution Gas Flow Rate.

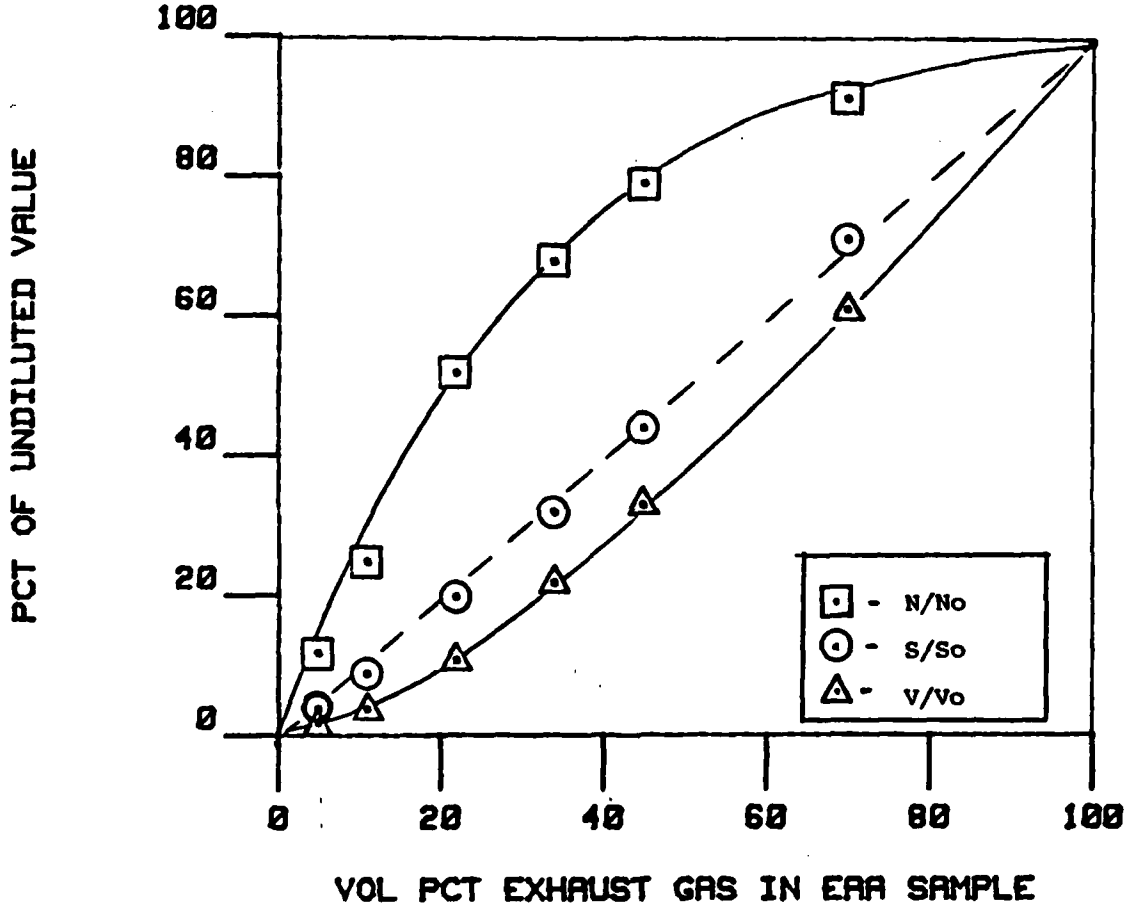


Figure 3. Dilution Study; The Variation of Various Measures of Soot Concentration with Volume Percent Exhaust Gas. N - Butane at ϕ = 0.1, v = 7.5 m/s.

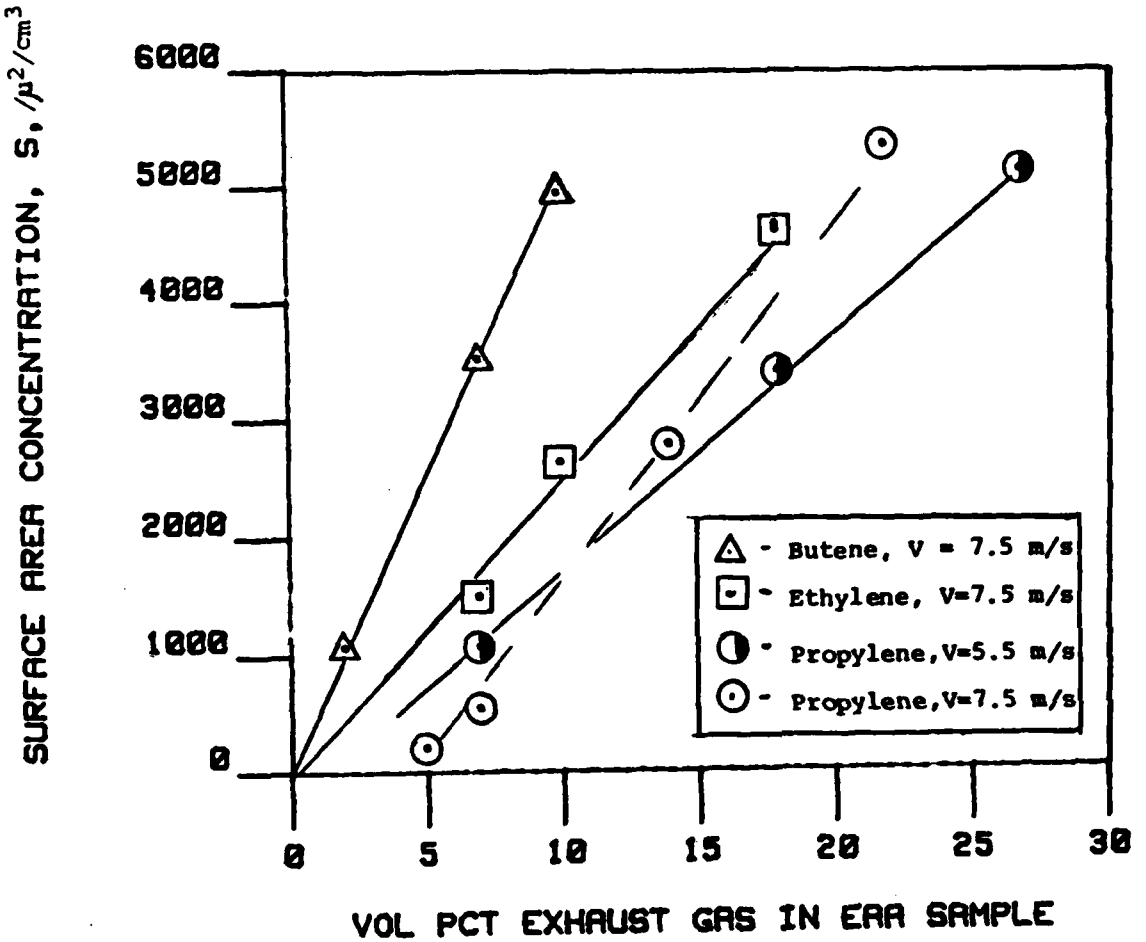


Figure 4. Variation of Aerosol Concentration (S, in μ^2/cm^3) with Volume Percent Exhaust Gas in the EAA Sample.

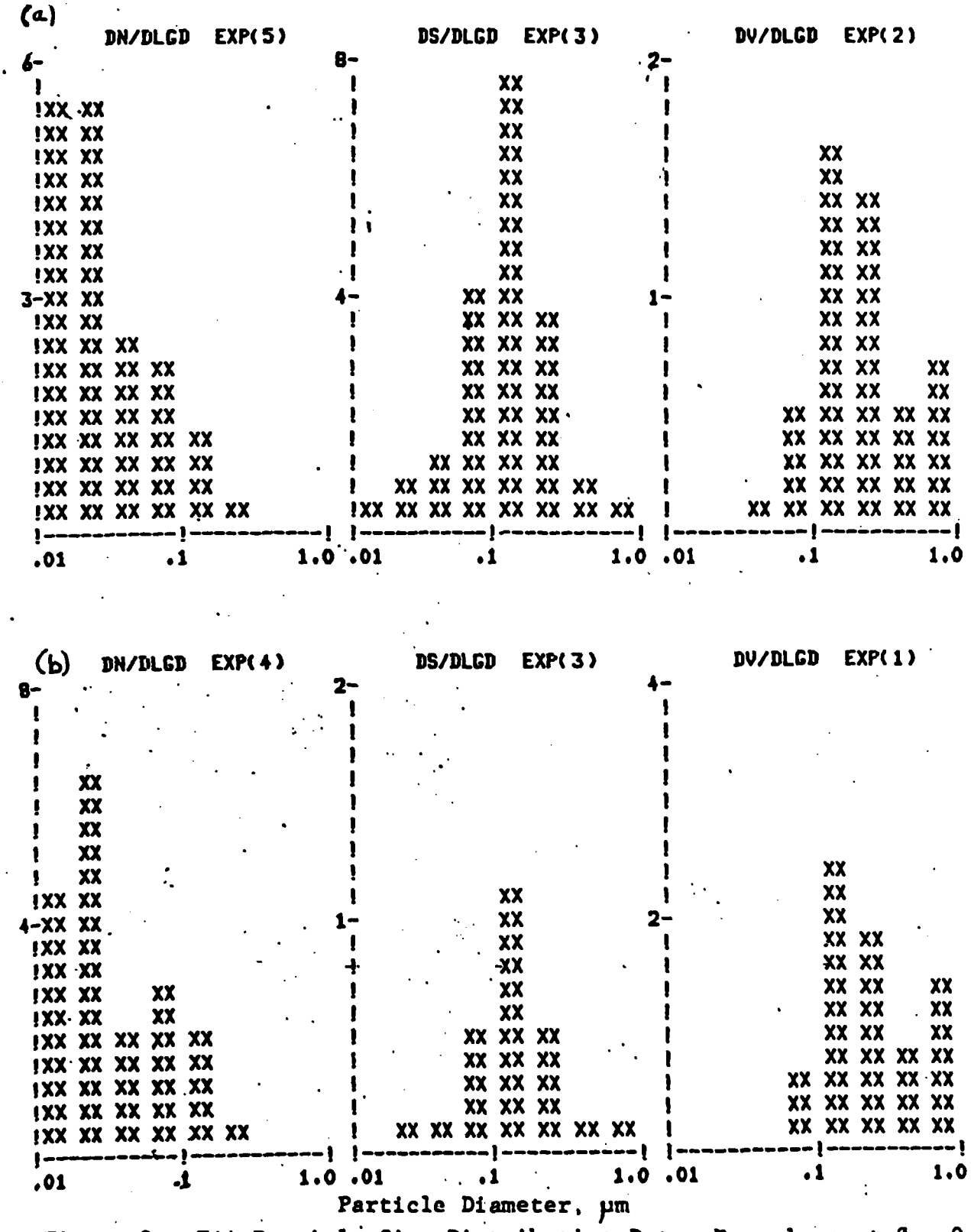


Figure 5. EAA Particle Size Distribution Data, Propylene at $\emptyset$ = 0.1, V = 7.5 m/s: (a) at 82% dilution, (b) at 93% dilution.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

DEVELOPMENT OF A MATHEMATICAL FOUNDATION FOR CELLULAR IMAGE PROCESSING

Prepared by:	Dr Gerhard X Ritter
Academic Rank:	Associate Professor
Department and University:	Department of Mathematics University of Florida
Research Location:	Air Force Armament Laboratory Guided Weapons Division Electro-Optical Terminal Guidance Branch
USAF Research Colleague:	Lt Col Lawrence A Ankeney, PhD
Date:	July 23, 1982
Contract No:	F49620-82-C-0035

DEVELOPMENT OF A MATHEMATICAL FOUNDATION FOR CELLULAR IMAGE PROCESSING

by

Gerhard X Ritter

ABSTRACT

Recent increases in image processing and pattern recognition activities have resulted in a proliferation of notation and basic operations for describing image processing algorithms that are often identical. A rigorous mathematical structure for aiding in the design, development, optimization and testing of image processing and image classification algorithms is, therefore, highly desirable. This study develops an initial foundation of such a structure for cellular image processing based on the interaction between algebra and topology. Specifically, we show that image dilations, erosions and other types of cellular logic operations can be expressed entirely in the topological language of limit points, and open and closed sets. The novel concept presented is the creation and use of cellular topological spaces. Endowing cellular sets with a topology provides a vehicle for transforming image manipulations to the standard mathematical terminology used in topology. Because of this, it now seems possible to provide a standardized mathematical formulation for all geometric image processing concepts. As examples, we provide some neighborhood operations for filtering, data compression, and image matching. Some of these are novel while others date back to Minkowski's development of the geometry of numbers in 1897.

Acknowledgements

The author would like to express his appreciation and gratitude to the Air Force Systems Command, the Air Force Office of Scientific Research, and the Southeastern Center for Electrical Engineering Education for providing the opportunity to spend a very interesting and valuable summar at the Air Force Armament Laboratory, Eglin AFB FL. In particular, he would like to acknowledge the staff of the Electro-Optical Terminal Guidance Branch of the Guided Weapons Division. Their hospitality and support made this summer research a most pleasant experience. Special thanks and sincere gratitude are due Lt Col Lawrence Ankeney, PhD, and Mr Neal Urquhart for their continued encouragement, for suggesting this research topic, and for their personal friendship which was exhibited throughout the ups and downs of this research.

Finally, the author would like to acknowledge the assistance of Mrs Janice Broxton in the preparation of this and other manuscripts.

I. INTRODUCTION

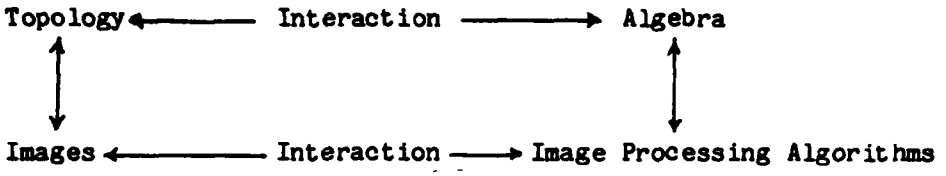
It is safe to say that modeling the brain constitutes the most difficult and most fascinating problem confronting modern science. One of the more intensely studied aspects of this problem concerns the visual systems. These studies are of special importance due to the need for rudimentary computer-based vision systems in industrial robots as well as autonomous guided weapons. In comparison to the eye, most of today's computer vision-pattern recognition systems lack real-time capabilities. This limitation is due to the sequential nature of the processes defining these systems. Because of this limitation, such systems are of little use in the guidance of high speed weapons. Cellular computers, on the other hand, have parallel processing capabilities and, thus, have the potential for evolving into highly sophisticated real-time image processing systems. In this sense, cellular computers are capable of behaving much more like the retinal and geniculate structure of the brain's vision system.

Although several types of cellular computers - or cyto computers - already exist, their full potential has not yet been realized. This is partially due to the non-existence of a unified, rigorous mathematical structure for cellular image processing. If such a structure existed - not just in terms of a uniform image processing language, but also as a mathematical environment for the testing, prediction of performance and extrapolation of new algorithms - then it would be feasible to test the optimal capabilities of existing as well as hypothetical cellular image processing systems. The main objective of this research is to create a mathematical foundation for such a structure. Once a solid foundation has been established, the structure can then be defined and built by extracting information from the store of knowledge comprising the foundation, very much like a growing tree extracts it nutrients from the soil it stands on.

Algorithms consist of algebraic manipulations of some sort while images consist of topological objects. Since image processing algorithms (algebraic operations) operate on images (topological objects), it seems natural to choose the rich mathematical theory of the interaction of

algebra with topology as a basis for a mathematical structure for image processing. Our approach is based on this rationale and is summarized in the following diagram:

Foundation for a Mathematical Structure for Image Processing



II. OBJECTIVES

The main objective of this research was the identification of a uniform mathematical structure for image processing. We did not attempt to attack the general problem in the short time span provided by the research contract. Instead, we focused our attention on exploring possible foundations for a uniform mathematical structure for image processing, and leaving the building of the structure for later investigations. Our specific objectives were:

(1) Investigate existing cellular image processing operations.

(2) Define the mathematics that can serve as a basis for a uniform mathematical theory of cellular image processing.

(3) Explore relationships between the mathematical basis and existing cellular operations.

(4) Compile a bibliography of books and articles concerned with or relating to cellular-image processing.

III. THE TOPOLOGICAL FOUNDATION OF CELLULAR IMAGE PROCESSING

The cellular automatas of von Neumann-Moore (1)-(5) and digital images share a common framework. Each image point can be viewed as a cell in a given state. By defining neighborhood relationships and cell transition functions on an image in terms of these neighborhood relationships, the configuration of the cell states in an image can be changed to a new configuration.

In this paper we shall consider transition functions that will change images in a predictable and useful fashion. Because of the dependence of these transition functions on the values of cells within the neighborhood of a range cell, these functions are also known as neighborhood transforms. Before describing neighborhood relations and operations, we shall define and give examples of topological spaces that provide the underlying geometric foundation for image processing.

Cellular automata and cellular image processing algorithms are based on the notion of neighborhood operations. This suggests that we study the topology of images by first choosing a subset of cells to serve as neighborhoods, and then carry over unaltered all other notions of topology. For this reason we choose the neighborhood definition for topological spaces.

Definition. A topological space is a set X together with a collection N of subsets of X called neighborhoods satisfying the following two axioms:

(a) Every point of X is in at least one neighborhood.

(b) If U and V are neighborhoods and $p \in U \cap V$, then there is a neighborhood W such that $p \in W \subset U \cap V$. Thus, we distinguish between neighborhoods and open sets (union of neighborhoods) since not every open set is an element of N.

Henceforth, R and Z will denote the set of real numbers and integers, respectively. Let Z^k denote the Cartesian product of k-copies of Z. With each $p=(p_1,\ldots,p_k) \in Z^k$ we associate the cell with center p, c(p) defined by

$$c(p) = \{q \mid q=(q_1,q_2,\ldots,q_k),\ q_i \in R \text{ and } |p_i-q_i| \leq 1/2 \text{ for } i=1,\ldots,k\}$$

The set of all k-dimensinal cells is denoted by C^k and is defined as $C^k = \{c(p) \mid p \in Z^k\}$. Since Z^k and C^k are in one-to-one correspondence, it is often customary to use p for the set c(p).

Definition. A cellular topological space consists of the set C^k together with a collection of neighborhoods satisfying axioms (a) and (b).

We now give examples of cellular spaces based on neighborhood configurations most frequently used in present day neighborhood image processing.

III.1 The von Neumann Topology

Let $L = \{-1,0,1\}$ and $p = (p_1,p_2,\ldots,p_k)$. The neighborhood N(c(p)) - or simply N(p) - of $c(p) \in C^k$ is defined by

$$N(p) = \begin{cases} \{c(p)\} & \text{if } \sum_{i=1}^{k} p_i \text{ is odd} \\ \{c(p_1,\ldots,p_{i-1},p_i+n,\ p_{i+1},\ldots,p_k) \mid 1 \leq i \leq k \text{ and } n \in L\} & \text{otherwise} \end{cases}$$

The collection $N = \{N(p) \mid p \in Z^k\}$ satisfies the neighborhood axioms. the topology thus derived is called the von Neumann topology for C^k.

Suppose $p=(i,j) \in Z^2$ and $q=(i,j,k) \in Z^3$. Then the possible neighborhoods of c(p) and c(q) are shown in Figure 1(a) and (b), respectively. The cruciform neighborhood shape in Figure 1(a) is due to von Neumann (10), hence the name for this topology.

Definition. A cell (point) in the von Neumann space is called even if the sum of its coordinates is even, otherwise it is called odd.

We note that if c(p) is odd, then the closure of N(p) is a cruciform neighborhood, and if c(p) is even, then $\{c(p)\}$ is closed.

It is also not difficult to show that in the von Neumann topology, sets are connected if and only if they are path-connected. However, we shall not use this fact and therefore omit its proof. As an example, the set $A \subset C^2$ shown in Figure 2(a) is connected and hence path-connected. On the other hand, the configuration $C = A \cup B$ shown in Figure 2(b) is not connected since $A \cap Cl(B) = Cl(A) \cap B = \emptyset$, where Cl denotes closure. This separation due to diagonally adjacent cells has resulted in numerous investigations (cf.(6),(7),(8) and (9).

III.2 The Moore Topology

The Moore topology for C^k is obtained by defining the neighborhood of a cell c(p), $p = (p_1,\ldots,p_k)$, as follows:

$$N(p) = \begin{cases} \{c(p_1+l_1, p_2+l_2, \ldots, p_k+l_k) \mid \ l_i \in L\} & \text{if } x_i \text{ is even for each } i \\ \{c(p)\} & \text{if } p_i \text{ is odd for at least one } i \end{cases}$$

As before, the collection $N = \{N(p) \mid p \in Z^k\}$ satisfies the neighborhood axioms. The topology generated by this neighborhood system is called the Moore topology.

The two possible neighborhood configurations of this system for C^2 are shown in Figure 3 where the shaded cell represents c(p).

In a Moore space, a point is <u>even</u> if each of its coordinates are even, otherwise it is called <u>odd.</u> Implications of using the neighborhood configuration N(p) (p even) for cellular automata were first investigated by Moore, (1) and (5).

Note that in the Moore topology, diagonally adjacent cells are connected.

III.3 <u>The 2^k - Topology</u>

Thus far every topology for C^k was generated by two distinct neighborhood configurations. We now define a more complex topology which is based upon a greater variety of neighborhood configurations.

The set C^1 becomes a topological space if for each $p \in Z$ we define neighborhood in C^1 by

$$N(p) = \begin{cases} \{c(p-1), c(p), c(p+1)\} \text{ if } p \text{ is even,} \\ \{c(p)\} \text{ if } p \text{ is odd} \end{cases}$$

The topology generated is called the <u>two-neighborhood topology for C^1.</u>

For $k \leq 1$ we define the 2^k-neighborhood topology (or simply 2^k-topology) for C^k as the product topology of the two-neighborhood space C^1. The four different neighborhood configurations for the 2^2-space C^2 are shown in Figure 4. The shaded cell represents the cell c(i,j).

IV. NEIGHBORHOOD TRANSFORMS

If $x = c(p)$ and $y = c(q)$ are two points in a cellular space, then their sum is the point $x+y$ defined by $x+y = c(p+q)$. The difference $x-y$ is defined by $x-y = c(p-q)$, and if $k \in Z$, then the scalar product kx is the point $kx = c(kp)$.

Definition. Let $X = C^k$ be a cellular space and $W = \{c(p_1,p_2,\ldots p_k) \in X \mid p_i \in L\}$.
The window at $x \in X$ is defined as $W(x) = \{x+y \mid y \in W\}$. The neighborhood system associated with X is said to satisfy the window condition if it satisfies the following two properties.

$$N(x) \subset W(x) \text{ for every } x \in X, \tag{1}$$

$$\text{if } z \in N(x) \cap N(y), \text{ then } x \in W(z) \text{ and } y \in W(z) \tag{2}$$

Hereafter we shall always assume that X is a cellular space whose neighborhood system satisfies the window condition. Note that the discrete space and our examples of cellular spaces satisfy the window condition.

Notation. The power set of X will be denoted by P(X). If $A \subset X$, then Int(A) denotes the interior of A, Cl(A) the closure of A, and C(A) the complement of A.

Definition. An image is an element of P(X) and the complement of the image is its backbround. An image transformation on X is a function $P(X) \longrightarrow P(X)$.

Definitions of a set in terms of a given set can be viewed as image transformations. Thus Int, Cl and C can all be thought of as image transformations. For instance, $Int: P(X) \longrightarrow P(X)$ is the transformation defined by $Int: A \longrightarrow Int(A)$.

Composition of image transformations will be denoted by $\cdot$.
In particular, $Int \cdot Cl(A) = Int(Cl(A))$. If $T_1,\ldots,T_n$ are image transformations, then the composition $T_n \cdot T_{n-1} \cdot \ldots \cdot T_1$ will be denoted by $\prod_{i=1}^{n} T_i$. If all the T_i's represent the same transformation T, then for $n \geq o$ we define $T^n = \prod_{i=o}^{n} T_i$ with $T^o(A) = A$.

Definition. A window configuration w is a function $w:X \longrightarrow P(X)$ such that $x \in w(x) \subset W(x)$.

If $B \subset X$, let $I(w,B)$ denote the set $I(w,B) = \{w(x) \mid w(x) \cap B \neq \emptyset\}$. A window relation for $A \subset X$ is a set $R(w,A)$ obtainable from $I(w,A)$ and/or $I(w,C(A))$ by intersection, union or complementation. Thus, $C(I(w,C(A))) = \{w(x) \mid w(x) \subset A\}$ and $C(I(w,A)) = \{w(x) \mid w(x) \cap A = \emptyset\}$ are two examples of window relations for A.

A local operator or window operator is a function $L:I(w,x) \longrightarrow P(X)$ such that $\emptyset \neq L(w(x)) \subset W(x)$.

Definition. An elementary neighborhood transform T is an image transformatin defined locally in terms of a window condiguration w, a window relation $R(w,\cdot)$ and a local operator L. Explicitely, $T(A) = \{ b \mid b \in L(w(x)), w(x) \in R(w,A)\}$.

A neighborhood transform is an image transformation that can be expressed as a composition of elementary neighborhood transforms.

The importance of neighborhood transforms in image processing is their potential for simultaneous (parallel) application to the neighborhood of every cell of an image. More specifically, elementary neighborhood transforms are the only image transformations that can be performed by a cellular image processing computer with window W as one-step transformations in parallel.

Examples. Some obvious examples of elementary neighborhood transforms are Int, Cl and C. For example, if $L(w(x)) = w(x) = N(x)$, then $T(A) = \{b \mid b \in L(w(x)), w(x) \in C(I(w,C(A))\} = \{b \mid b \in N(x), N(x) \subset A\} =$ Int (A). Similarly, let $w(x) = N(x)$ and $L(w(x)) = \{x\}$. Then $Cl(A) = \{b \mid b \in L(w(x)), w(x) \in I(w,A)\}$.

The transform $G(A) = \{b \mid b \in N(a), a \in A\}$ is an elementary neighborhood transform. In order to verify this, let $w(x) = \{x\}$, $L(w(x)) = N(x)$ and $R(w,A) = I(w,A)$. Note that $G(A)$ is the smallest open set containing A.

The transform $F=C \cdot G \cdot C$ is an example of a neighborhood transform that transforms A into the largest closed set contained in A. In particular, we have the relationship

$$F(A) \subset A \subset G(A). \tag{3}$$

We conclude this section by defining two fundamental neighborhood transforms.

Definition. The dilation transform D is the image transformation defined by $D(A) = G(A) \cup Cl(A)$. The shrinking transform is the image transformation S defined by $S = C \cdot D \cdot C$.

Theorem 1. The transforms D and S are elementary neighborhood transforms.

Proof. (1) Let $w(x) = \{x\}$, $L(w(x)) = N(x) \cup \{y \mid x \in N(x) \cap N(y)\}$ and $T(A) = \{b \mid b \in L(w(x)), w(x) \in I(w,A)\}$. Since $N(x) \subset G(A)$ and $\{y \mid x \in N(x) \cap N(y)\} \subset Cl(A)$ whenever $w(x) \in I(w,A)$, $T(A) \subset D(A)$.

In order to show that $D(A) \subset T(A)$, let $b \in D(A)$. Then either $b \in G(A)$ or $b \in CL(A)$ or both. If $b \in G(A)$, then $b \in N(x)$ for some $x \in A$. But $N(x) \subset L(w(x)) \subset T(A)$ for every $x \in A$. Hence $b \in T(A)$. If $b \in Cl(A)$, then $N(b) \cap A \neq \emptyset$. Hence there is an $x \in N(b) \cap A$ and therefore, $x \in N(b) \cap N(x)$ with $\{x\} = w(x) \in I(w,A)$. Thus, $b \in \{y \mid x \in N(y) \cap N(x)\} \subset T(A)$.

(2) To prove that S is an elementary neighborhood transform, reverse the roles of w, L and R by letting $w(x) = N(x) \cup \{y \mid x \in N(x) \cap N(y)\}$, $L(w(x)) = \{x\}$ and $T(A) = \{b \mid b \in L(w(x)), w(x) \in C(I(w,C(A)))\}$. The argument for showing that $T(A) = S(A)$ is now similar to the argument in part (1).

Theorem 2. $a \in D(b)$ if and only if $b \in D(a)$.

Proof. If $a = b$, there is nothing to prove. Suppose $b \neq a \in D(b)$. Then either $a \in G(b)$ or $a \in Cl(b)$. If $a \in G(b)$, then $a \in N(b)$. Hence $b \in Cl(a) \subset D(a)$. If $a \in Cl(b)$, then $b \in N(a) \subset D(a)$. Thus $a \in D(b)$ implies that $b \in D(a)$. The proof of the converse is analagous.

The next theorem is a direct consequence of the definition of D and S.

Theorem 3. Let $A \subset X$. Then:

(1) $S(A) = F(A) \cap Int(A)$

(2) $Int(A) \subset S(A) \subset A \subset Cl(A) \subset D(A)$

(3) If A is open (closed), the D(A) and S(A) are closed (open)

Example. Let $X = C^2$ have the von Neumann topology and let $A \subset X$ be the shaded region shown in Figure 5(a). The region A represents a digitized version of a tank. The "holes" (unshaded cells) in the tank and the isolated shaded cell represent "noise" (artifacts) introduced by the digitization process.

Figures 5(b) and (c) show the effects of applying G and D to A, respectively. Note that D(A) is simply connected (both as a subset of E^2 and as a subset of C^2.). Roughly speaking, D removes small holes and produces a "smoother" looking version of A. On the negative side, the shaded isolated cell has been dilated and the cavity beneath the gun has disappeared.

The shrinking transform S has just the opposite effect (Figure 5(d)). The isolated cell has disappeared and cavities have been enlarged. Note also that the gun, which represents a cavity in C(A), has vanished.

V. THE ALGEBRAIC FOUNDATION OF CELLUAR IMAGE PROCESSING

All cellular logic or neighborhood operations reviewed in the appended list of references can be expressed in terms of the Minkowski algebra of sets. Furthermore, there is a direct link between cellular topologies and the Minkowski algebra that allows the reformulation of the algebraic operations in terms of topological operations. Thus, the interaction between cellular topology and the Minkowski algebra is a mathematical duality. It is for these reasons that we chose the Minkowski algebra for representing cellular image processing manipulations.

A thorough discussion of Minkowski's algebra can be found in (4). It includes such manipulations as addition, subtraction, scalar and vectorial multiplication, complementation, union and intersection. Because of our definition of neighborhood transforms, we also include such operations as local averaging of sets, local reflections and local spatial transformations. However, for reasons of brevity we shall only state the fundamental concepts of addition and subtraction.

Definition. Let A and B be subsets of E^k. The Minowski sum of A and B is defined vectorially as $A \times B = \{a+b \mid a \in A, b \in B\}$. The Minkowski subtraction of B from A is defined in terms of complementation by $A/B = C(C(A) \times (-B))$, where $-B = \{b \mid -b \in B\}$.

The connection between Minowski's algebra and neighborhood transforms is implicitely contained in the specifications of L(w) and R(w,·). For instance, I(N(x),A) corresponds to AxN(o) while C(I(N(x),C(A))) corresponds to A/N(o). As a particular example of this relationship we state and prove the following:

Theorem 4. If X is a von Neumann space, then D(A) = AxN(o) and S(A) = A/N(o).

Proof. Let $w(x) = \{p \mid |p - x| \leq 1\}$. If $a \in X$ is even, then w(a) = N(A) and if a is odd, then w(a) = Cl(N(a)). Thus, $\bigcup_{a \in A} w(a) = G(A) \cup Cl(A) = D(A)$. On the other hand, $A \times N(o) = \{a+b \mid b \in N(o), a \in A\} =$

$$\{p \mid p \in N(o) \times \{a\}, a \in A\} = \{p \mid p \in w(a), a \in A\} = \bigcup_{a \in A} w(a).$$

Therefore, AxN(o).

Using this fact and complementation, we obtain $A/N(o) = C(C(A)xN(0) = C(D(C(A))) = S(A)$. Q.E.D.

Obviously, similar results can be obtained for other cellular spaces.

VI. TOPOLOGICAL FILTERS

As we observed in the last section, the transform D smoothens images, removes small holes, but enlarges the image and isolated cells beyond their actual size. On the other extreme, S erodes an image and creates large cavities. Thus it becomes natural to consider compositions of these transforms in order to obtain an intermediate effect of these two extremes.

Definition. If $A \subset X$, then the exterior hull of A is denoted by E(A) and is defined by $E(A) = S \cdot D(A)$. The interior hull I(A) of A is defined as $I(A) = D \cdot S(A)$.

Example. Figure 6(a) shows the exterior hull of A, where A is the image shown in Figure 5(a). In terms of the algebraic manipulation, we have $E(A) = (AxN(o))/N(o)$. Note that E(A) is about the same size as A and that the isolated cell has not been enlarged while the holes have disappeared. Thus E might be viewed as a noise filter since it removes small concavities and holes.

Similar observations can be made for the interior hull $I(A) = (A/N(o))xN(o)$, as shown in Figure 6(b). Even though cavities still appear enlarged, erosion is not as severe as for S(A). Observe also that the front and rear end of the tank appear more rounded.

An important and interesting problem is the classification of images that remain invariant under neighborhood transforms. Before considering images that remain invariant under the transforms E and I, we recall the definition of regular sets.

Definition. An open set A is called regular if A = Int(Cl(A)). A closed set A is called regular if A = Cl(Int(A)).

The properties of regular sets summarized in the next theorem are well known (cf (2)).

Theorem 5.

(1) If A is closed, then Int(A) is regular.

(2) If A is open, the Cl(A) is regular.

(3) If A is open (closed) and regular, then Cl(A) is closed (open) and regular.

(4) If A and B are open and regular sets, then $A \subset B$ if and only if $CL(A) \subset Cl(B)$.

(5) If A and B are closed and regular sets, then $A \subset B$ if and only if $Int(A) \subset Int(B)$.

(6) If A and B are closed and regular sets, then $A \cup B$ is a closed and regular set.

(7) If A and B are open and regular sets, then $A \cap B$ is an open and regular set.

Theorem 6. If A is open and regular, then E(A) = A. If A is closed and regular, then I(A) = A.

Proof. If A is open, then D(A) = Cl(A). But then $E(A) = S \cdot D(A) = Int(Cl(A))$ and the result follows since A is regular.

If A is closed, the S(A) = Int(A). Hence $I(A) = D \cdot S(A) = Cl(Int(A))$. Again by regularily, Cl(Int(A)) = A. This proves Theorem 6.

By combining Theorems 5 and 6 further results may be obtained. For instance, if A is closed, then Int(A) remains invariant under E. Similarily, if A is open, then the interior hull of Cl(A) is Cl(A).

The proof of Theorem 6 in conjunction with the next theorem show that E and I map open (closed) sets to open (closed) sets.

Theorem 7. If A is closed, the $E(A) = F(G(A))$ and if A is open, the $I(A) = G(F(A))$.

Proof. If A is closed, then $D(A) = G(A)$. Hence $E(A) = S\ D(A) = F(G(A))$. Similarily, if A is open, then $S(A) = F(A)$ and $I(A) = D \cdot S(A) = G(F(A))$. This proves Theorem 7.

Theorem 8. $S(A) \subset I(A) \subset A \subset E(A) \subset D(A)$.

Proof. $E(A) = S \cdot D(A) \subset D(A)$ by Theorem 3. Now let $a \in A$. Then $N(a) \subset D(A)$. Hence $a \in \{a\} = S(N(a)) \subset S \cdot D(A)$. Therefore, $A \subset E(A)$. The proof that $S(A) \subset I(A) \subset A$ is similar and therefore omitted. This establishes Theorem 8.

Theorem 9. $E^2(A) = E(A)$ and $I^2(A) = I(A)$.

Proof. Obviously $I^2(A) = I(I(A)) \subset I(A)$. On the other hand, $I \cdot I(A) = D(S \cdot D(S(A))) = D(E(S(A))) \supset D(S(A)) = I(A)$. A similar argument shows that $E^2(A) = A$. This establishes Theorem 9.

Theorem 10. $I(D(A)) = D(A)$ and $E(S(A)) = S(A)$.

Proof. Clearly, $I(D(A)) \subset D(A)$. But $I(D(A)) = D(S \cdot D(A)) = D(E(A)) \supset D(A)$. Hence $I(D(A)) = D(A)$.

Similarily, $S(A) \subset E(S(A))$ and $E(S(A)) = S(D \cdot S(A)) = S(I(A)) \subset S(A)$. Therefore $E(S(A)) = S(A)$. This proves Theorem 10.

VII. NEIGHBORHOOD REDUCTION AND RECONSTRUCTION

Neighborhood reduction is a data compression device. This technique, defined in terms of neighborhood transforms, reduces the image to a few cells. Reacquisition of the original image from the reduced image can be accomplished via neighborhood reconstruction. Thus, neighborhood reduction compresses data without loss of information.

Definition. Let i be a non-negative integer and $A \subset X$. The i-th reduction of A is defined as $R_i(A) = S^i(A) \cap C \cdot I \cdot S^i(A)$. The neighborhood reduction of A is the disjoint union $R(A) = \cup R_i(A)$.

Since $S^i(A) = \{ a \in A \mid D^i(a) \subset A \}$ and $R_i(A) = S^i(A) \cap C(D(S^{i+1}(A)))$, we can express $R_i(A)$ in set theoretic terms as $R_i(A) = \{ a \mid D^i(a) \subset A$ and $D^{i+1}(b) \not\subset A$ whenever $b \in D(a) \}$. This last expression provides a clearer view of the geometric manipulations that are required in order to obtain $R_i(A)$. The next example should elucidate the importance of neighborhood reductions.

Example. Let A be the image shown in Figure 6(a) with the isolated shaded cell deleted. Figure 7 shows A (dotted outline) superimposed on R(A). In order to show the position of $R_i(A)$ with respect to A, we labeled the cells of $R_i(A)$ by i (i = 0,1,2,3). Note that $R_i(A) = \emptyset$ for $i > 3$.

An important observation is that the reduction R(A) consists of only 44 cells. This is less than 1/3 of the 159 cells composing the image A. Another useful observatin is that R(A) can be viewed as a weighted skeleton of A, where a cell of the skeleton has weight i if it belongs to $R_i(A)$. As can be inferred from the reconstruction procedure, the largest numbered cells correspond to the bulk of the image while the cells with small weights correspond to the finer features. In tracking an object such as a moving target, it is often sufficient to only track cells with large weights, thus allowing further data reduction.

Definition. The interior radius $m(A)$ of A is defined as $m(A) = \max \{i \mid D^i(a) \subset A, a \in A\}$ if the maximum exists, otherwise $m(A) = \infty$.

The exterior radius $M(A)$ of A is defined as $M(A) = \min \{i \mid A \subset D^i(a), a \in A\}$ if the minimum exists, otherwise $M(A) = \infty$.

It follows that $0 \leq m(A) \leq M(A)$. Also, $M(A) < \infty$ if and only if A is finite. However it is possible that $m(A) < \infty$ even though $M(A) = \infty$. Unless otherwise stated, we shall always assume that $m(A) < \infty$. This complies with the spirit of image processing by computer where images are always finite.

Definition. Let $A \subset X$ and $0 \leq k \leq m(A)$. The k-th partial reconstruction r of A from $R(A)$ is defined as $r_k(A) = \bigcup_{i=m(A)-k}^{m(A)} D^i(R_i(A))$.

Theorem 11. $r_{m(A)}(A) = A$.

Proof. If $p \in R_i(A)$, then $D^i(p) \subset A$. Hence $D^i(R_i(A)) \subset A$ for each $i \geq 0$ and, therefore, $r_{m(A)}(A) \subset A$.

In order to show that $A \subset r_{m(A)}(A)$, let $A_i = \{a \mid \exists\ b \in A$ with $a \in D^i(b) \subset A\}$. Since $A_0 = A$, $A = \bigcup A_i$. Now let $a \in A$ and $k = \max \{i \mid a \in A_i\}$. Thus $a \in D^k(b) \subset A$ for some $b \in A$. Suppose there is a $c \in D(b)$ such that $D^{k+1}(c) \subset A$. By Theorem 2, $b \in D(c)$. But then $D^k(b) \subset D^{k+1}(c)$. Therefore $a \in D^{k+1}(c) \subset A$ and $a \in A_{k+1}$, contradicting the fact that k was maximal with respect to this property. Hence $b \in R_k(A)$ and $a \in D^k(B) \subset D^k(R_k(A)) \subset r_{m(A)}(A)$. This proves Theorem 11.

It now follows that the definition of r_k provides an algorithmic method of reconstructing A- or specific subsets of A- knowing only $R(A)$. As an example, Figure 8 shows the first partial reconstruction $r_1(A)$ of A, where $R(A)$ is as in the previous example.

VIII. NEIGHBORHOOD METRICS

A neighborhood metric is a metric between two images A and B that can be computed in terms of neighborhood transforms. The metric we are about to define was first proposed by Minkowski (3)-(4). Ours will be a discrete version of this metric.

Definition. The Minkowski distance between two images A and B in a cellular space is defined as $d(A,B) = \min\{k \mid A \subset D^k(B) \text{ and } B \subset D^k(A)\}$.

Let $A^k = D^k(A)$, $N = N(O)$ and X the von Neumann space. It then follows from Theorem 4 that $d(A,B) = \min\{k \mid A \subset B \times N^k, B \subset A \times N^k\}$. Thus, the distance between A and B can be obtained by successive dilation of the neighborhoods of the points of A and B.

To avoid frivolity, we shall assume that X is a non-discrete space. We continue to assume that X satisfies the window condition.

Theorem 12. If $Y \subset P(X)$ consists of finite images, then d is a metric on Y.

Proof. (1) Clearly $d(A,B) \geq 0$. If A = B, then $A \subset B = D^0(B)$ and $B \subset A = D^0(A)$. Therefore d(A,B) = 0. Conversely, if d(A,B) = 0, then $A \subset D^0(B) = B$ and $B \subset D^0(A) = A$. Thus A = B.

(2) d(A,B) = d(B,A) is obvious.

(3) Let d(A,B) = n and d(B,C) = m. Then $A \subset D^n(B)$ and $B \subset D^m(C)$. Hence, $A \subset D^n(D^m(C)) = D^{n+m}(C)$. Similarily, $B \subset D^n(k)$ and $C \subset D^n(B)$. therefore $C \subset D^m(D^n(A)) = D^{m+n}(A)$ and $d(A,C) \leq n+m = d(A,B) + d(B,C)$. This proves Theorem 12.

The Minkowski distance has some interesting properties with respect to unions, dilations and Minkowski addition. We summarize some of these in the next theorem.

Theorem 13. Let A,B,C and D be finite images in X. Then:

(1) $d(A \cup B, C \cup D) \leq \max \{ d(A,C), D(B,D) \}$.

(2) $d(A \cup B, A) \leq d(A,B)$.

(3) $d(A \cup B, C) \leq d(A,B) + d(B,C)$.

(4) $d(A \times B, C \times D) \leq d(A,C) + d(B,D)$.

(5) $d(A,B) = d(A^k, B^k)$.

Since the proofs of these properties are all very similar, we shall only verify property (1). Let $m = d(A,C)$, $n = d(B,D)$ and $k = \max\{m,n\}$. Then $A \subset C^m$, $B \subset D^n$ and therefore $A \cup B \subset C^m \cup D^m \subset (C \cup D)^k$. Similarily, $C \cup D \subset (A \cup B)^k$. Thus $d(A \cup B, C \cup D) \leq k$.

The Minkowski metric provides a useful measure of similarity between image. As such it can be used as an image matching device. Its definition in terms of the dilation transform D reflects the geometrical structure of images. In comparison to many other similarity measures it is also less sensitive to small pertubations caused by noise. Because of its dependence on the elementary neighborhood transform D, the question arises as to what relationships exist between d and neighborhood transformations involving D in their factorization.

Large pertubations affect the Minkowski metric severely. d(A,B) may be large even though A and B represent the same configuration. This is the case when A occupies a different location than B. We provide a partial remedy for this situation by defining a translation invariant neighborhood distance based on the Minkowski metric.

Let $v(p) = v(p_1,\ldots,p_k) \in X$, $V = \{ v_i(p) \mid p_i = 1, p_j = 0 \text{ if } j \neq i \}$ and, to simplify notation, let $v_i = v_i(p)$. Note that $V \subset W$. For each $v \in X$, we may define an image translation T_v by $T_v(A) = A \times \{v\}$, where x denotes Minkowski addition. Specifically, $T_v(A)$ is a translation of the image A by the "vector" v.

As a first result we have that for each $v_i \in V$, T_{v_i} is an elementary neighborhood transform. To ascertain this claim, let $w(x) = \{ x \}$ and $L(w(x)) = \{ x + v_i \}$. Then $T_{v_i}(A) = \{ b \mid b \in L(w(x)), w(x) \in I(w,A) \}$. It now follows that for any $n \in Z$, T_{nv_i} is a neighborhood transform since $T_{nv_i} = \prod_{j=0}^{n} T_{v_i}$. More generally, since any $v \in X$ can be expressed as $v = \sum_{i=1}^{k} n_i v_i$, with $n_i \in Z$, $T_v = \prod_{i=1}^{k} T_{n_i v_i}$. Hence T_v is a neighborhood transform for any $v \in X$.

Using these translation neighborhood transforms we are now able to define a translation invariant similarity measure between images.

Definition. The Minkowski similarity s for images is defined by $s(A,B) = \min \{ d(A,T_v(B)) \mid v \in X \}$, where d denotes the Minkowski distance.

Since $s(A,B) \leq d(A,B)$; s is a more discriminant measure of image similarity. Also, since $s(A,T_v(A)) = 0$ for each $v \in X$, s is translation invariant. In particular, s is a metric on $Y = \{ [A] \mid A \subset X \text{ is finite} \}$, where $[A] = \{ B \mid B \subset X, T_v(B) = A \}$.

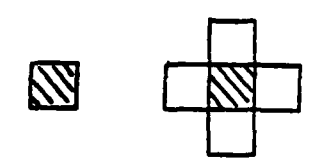
The Minkowski similarity solves the problem caused by translations but introduces the problem of finding an optimal algorithm that determines v such that $d(A,T_v(B))$ is minimal. Furthermore, it is not the ultimate matching device but simply one of many neighborhood techniques derived from the interaction of the image geometry with the image algebra. It is the capabilities, optimization, advantages and disadvantages of these cellular techniques and their proper place in the total scheme of cellular image processing that needs to be investigated.

IX. RECOMMENDATIONS

It is now time to place our results into a proper practical perspective. As we have indicated, current neighborhood image processing operations can be expressed in terms of the Minkowski algebra of sets. The geometric dual of this algebra can be obtained from the topology of the image space, and this duality provides a rich foundation for a rigouous mathematical structure for cellular image processing. However, we have not attempted to build such a structure. In fact, we have barely tapped the resources of this foundation. We therefore recommend the development of a mathematical structure for cellular image processing based on the topologic-algebraic foundation proposed in this paper.

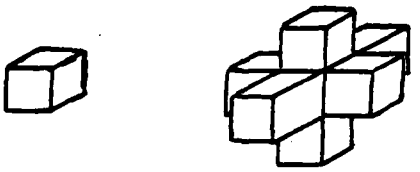
Since this research is still in the embryonic stage, various primary steps in the initial development of the structure can be taken. For instance, a valuable first step could be the cataloguing and transformation of existing neighborhood techniques, that aim at fulfilling the same or similar tasks, into the Minkowski-topologic domain. An excellent choice would be the multitude of neighborhood algorithms used for skeletonizing images. The geometric performance of these algorithms could then be tested via the theorems of topology and algebra instead of using specific examples which all to often result in false conclusions. The experience and insights gained from the primary steps can then be applied to further develop the structure. Therefore, ad astra.

THE von NEUMANN NEIGHBORHOODS



c(p) odd c(p) even

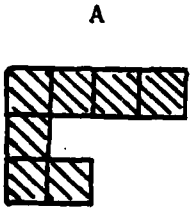
(a)



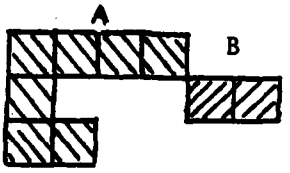
c(p) odd c(p) even

(b)

FIGURE 1



(a)



(b)

FIGURE 2

THE MOORE NEIGHBORHOODS



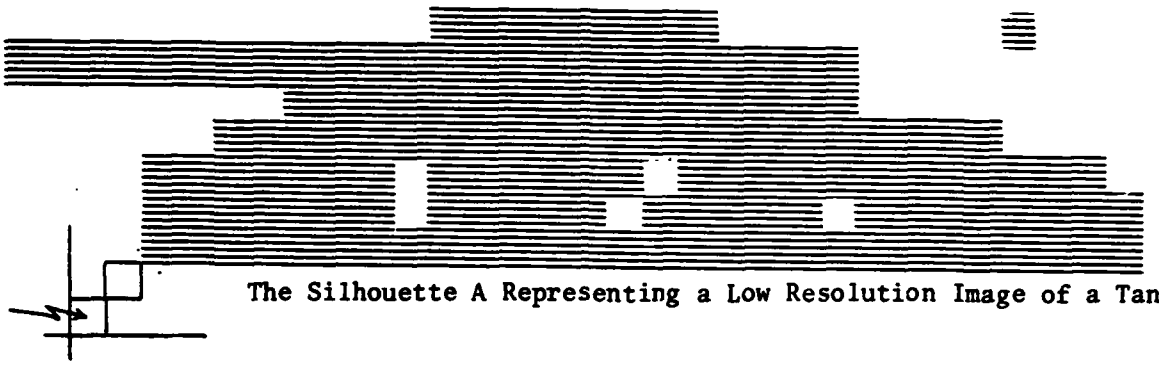
c(p) odd c(p) even

FIGURE 3



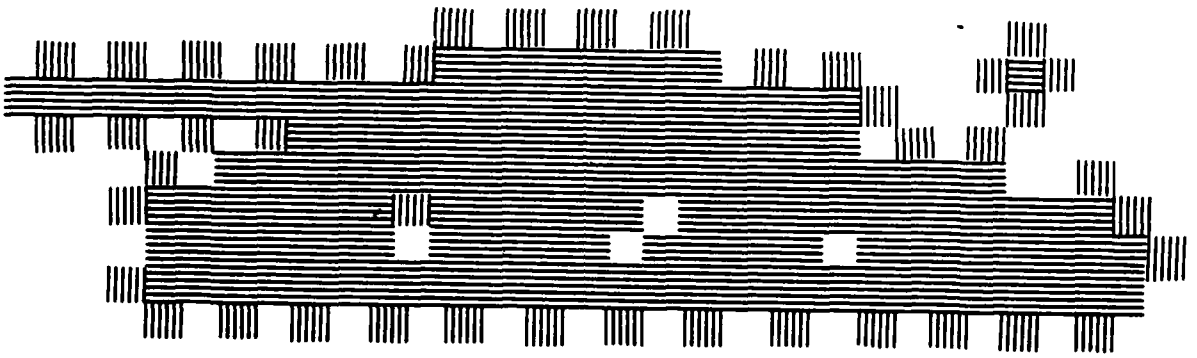
The 2^k Basic Neighborhoods if k = 2.

FIGURE 4



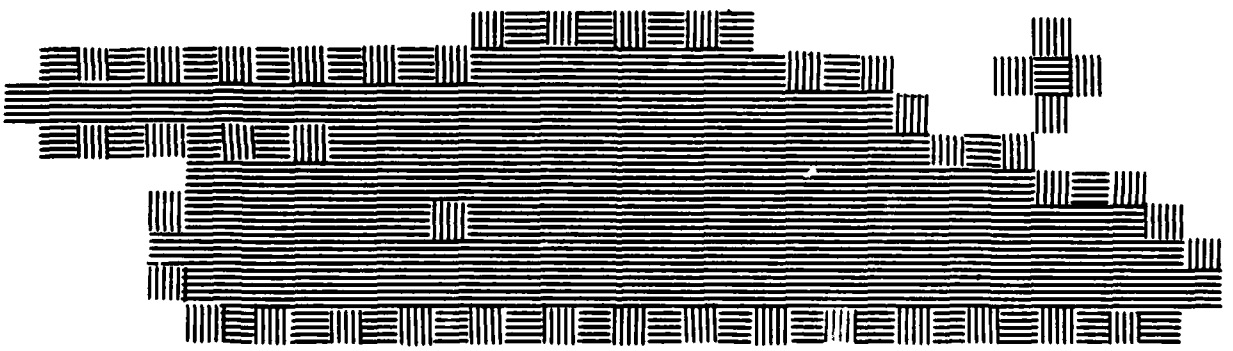
The Silhouette A Representing a Low Resolution Image of a Tank

FIGURE 5(a)



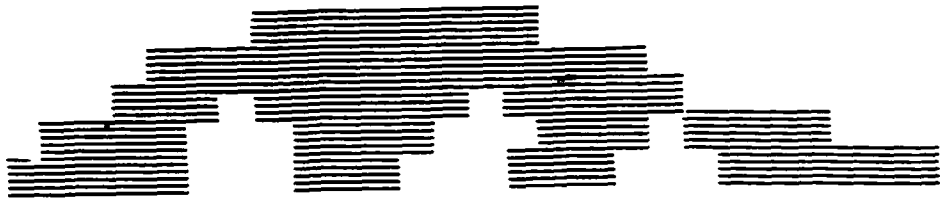
G(A)

FIGURE 5(b)



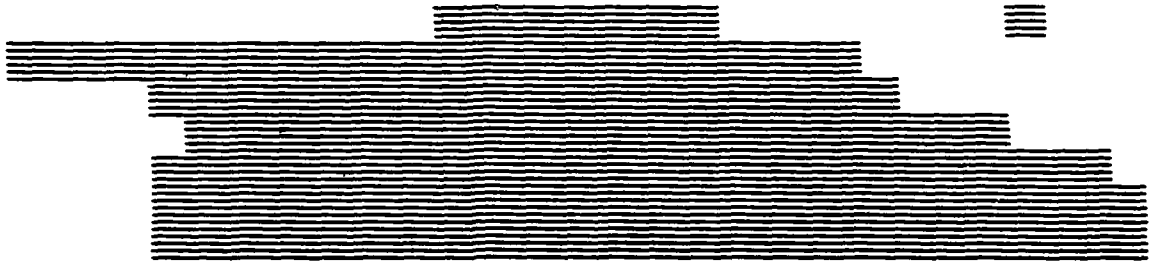
$D(A) = G(A) \cup C1(A)$

FIGURE 5(c)



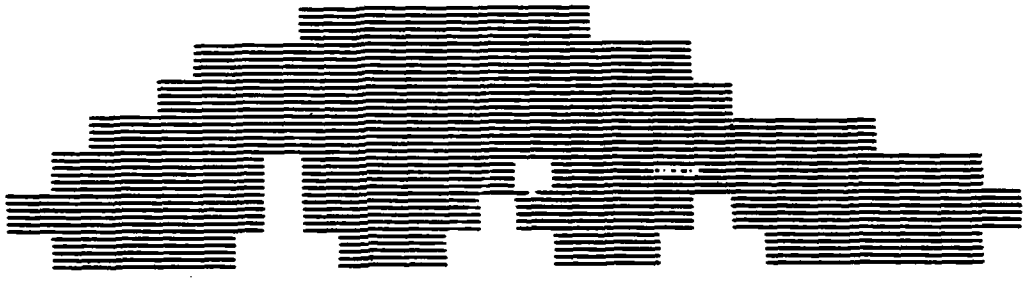
$S(A) = F(A) \cap Int(A)$

FIGURE 5(d)



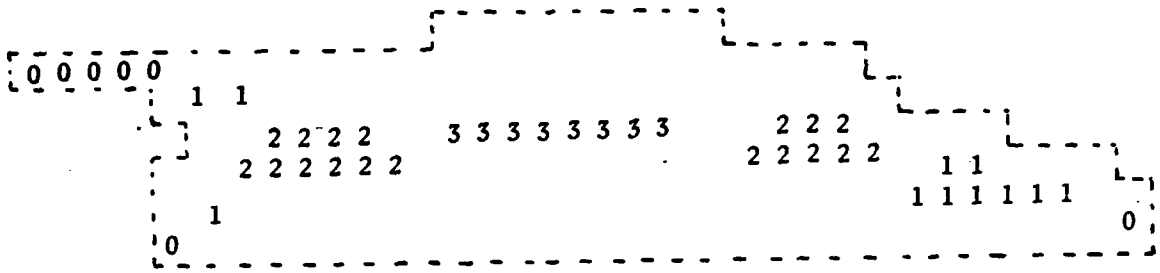
E(A)

FIGURE 6(a)



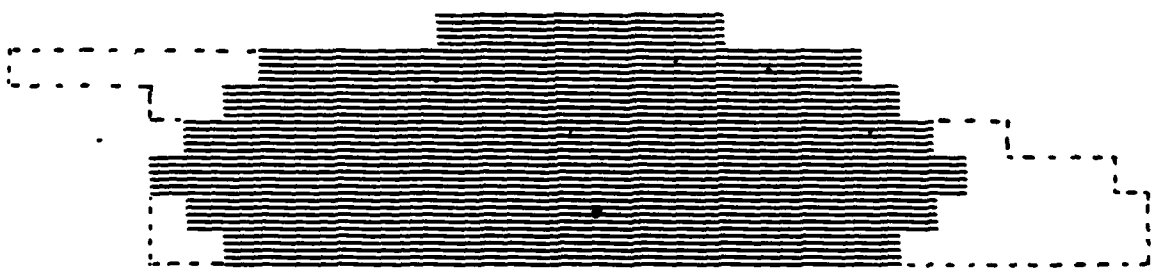
I(A)

FIGURE 6(b)



R(A)

FIGURE 7



$r_1(A)$

FIGURE 8

REFERENCES

1. A W Burks, Ed., Essays on Cellular Automata, University of Illinois Press, 1970.

2. J Dugundji, Topology, Allyn and Bacon Inc, Boston, 1967.

3. H Hadwiger, Vorlesungen uber Inhalt, Oberflache und Isoperimetrie, Springer Verlag, Berlin-Gottingen-Heidelberg, 1957.

4. H Minkowski, Gesammelte Abhandlungen, Teubner Verlag, Leipzig-Berlin, 1911.

5. E F Moore, "Machine Models of Self-Reproduction," AMS Proceedings of Symposia in Applied Math, 14, (1962), 17-33.

6. A Rosenfeld, "Connectivity in Digital Pictures," J Assoc Comp Mach, 17, (1970), 146-160.

7. _____, "Adjacency in Digital Pictures," Inf and Control, 26, (1974), 14-33.

8. _____, and A C Kak, Digital Image Processing, Academic Press, New York, 1976.

9. _____, "Digital Topology," MAA Monthly, 86, (1979), 621-630.

10. J von Neumann, Theory of Self-Reproducing Automata, Edited by A W Burks, University of Illinois Press, 1966.

X. APPENDIX

A SURVEY OF REFERENCES

RELEVENT TO

CELLULAR IMAGE PROCESSING

The following list of publications consists of articles and books either directly concerned with cellular image processing or with the notions and concepts used in the cellular image processing literature. For example, the undefined notion of a sigma algebra referred to in Matheron's papers can be found in Rudin's book on analysis. Similarly, various concepts from topology have been used by many authors. Hence, some readable elementary topology texts are included.

Since this list is neither complete nor up-to-date, I hope that subsequent readers are willing to aid in its expansion.

REFERENCES

1. N M Amosov, Modeling of Thinking and Mind, McMillan, London, 1967

2. Amoroso and H Yamada, "Structural and Behavioral Equivalences of Tesselation Automata," AROD Rep 6454-8-M, US Army, 1971

3. L A Ankeney, The Design and Performance Characteristics of a Cellular Logic 3-D Image Classification Processor, PhD Dissertation, Air Force Institute of Technology, Air Force University, Wright Patterson AFB OH, 1981

4. L A Ankeney and G X Ritter, "Applications of Cellular Topology in Image Processing," Preprint, 1982

5. M Arbib, The Metaphorical Brain: An Introduction to Cybernetics As Artificial Intelligence and Brain Theory, Wiley and Sons, New York, 1972

6. W T Beyer, "Recognition of Topological Invariants by Iterative Arrays," Project MAC, MIT Tech Rep, 1969

7. G Birkhoff, "Lattice Theory," AMS Colloquium Publ., Providence RI, 1948

8. W Blascke, Kreis und Kugel, Veit Co, Leipzig, 1911

9. H Blum, "A Transformation for Extracting New Descriptors of Shape," Models for the Perception of Speech and Visual Form, MIT Press, 1967

10. G S Brindley, Physiology of the Retina and Visual Pathways, Arnold, London, 1960

11. B Bullock, "The Performance of Edge Operators on Images with Texture," TR AD/A-006 230, Hughes Res Labs, Malibu, CA, 1978

12. A W Burks, Ed., Essays on Cellular Automata, Univ of Ill. Press, 1970

13. M Cannon, "Terrain Contour Matching, Sensitivity to Heading and Ground Speed Errors," AMRLTR-77-84, Wright Patterson AFB OH, 1978

14. E F Codd, Propagation, Computation, and Construction in Two-Dimensional Cellular Spaces, PhD Thesis, University of Michigan, 1965

15. E F Codd, Cellular Automata, Academic Press, New York-London, 1968

16. A J Cote, The Search for the Robots, Basic Books Inc, New York, 1967

17. H S M Coxeter, Regular Polytopes, McMillan, New York, 1963

18. P Danielson and S Levialdi, "Computer Architectures for Pictorial Information Systems," IEEE Trans on Comp., Nov 1981, 53-67

19. P Delfiner, "A Generalization of the Concept of Size," J of Microscopy, 95(2), (1972), 203-216

20. A Domon, "A Flexible Three Dimensional Cellular Space," Proc of the 3rd Int Congress of Cybernetics and Systems, Springer-Verlay Berlin, 1975

21. R O Duda and P E Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973

22. M J B Duff, "Cellular Logic and Its Significance in Pattern Recognition," AGARD Proc on Artificial Intelligence No. 49, 25, (1971), 1-13

23. M J B Duff, "A Cellular Logic Array for Image Processing," Pattern Recognition, 5, (1973), 229-247

24. M J B Duff, "CLIP4: A Large Scale Integrated Circuit Array Parallel Processor," Proc 3rd Int Conf on Pattern Recognition, (1976), 728-733

25. J Dugundji, Topology, Allyn and Bacon Inc, Boston, 1967

26. G B Elspas, "Topological Constraints on Interconnection-Limited Logic," IEEE Special Publ Switching Circ Theory and Logical Design, S-164, (1964), 133-147

27. B Elspas, et al, "Properties of Cellular Arrays for Logic and Storage," Final Report, Contract AF 19(628)-5828, SRI Proj 5876 Stanford Res Inst, 1967

28. B Elspas and J B Turner, "Theory of Cellular Logic Networks and Machines," TR AFCRL-68-0148, Stanford Res Inst and Air Force Cambridge Res Labs, 1968

29. S Ginzburg, An Introduction to Mathematical Machine Theory, Addison Wesley, Reading MA, 1962

30. S Ginzburg, Algebraic Theory of Automata, Academic Press, New York-London, 1968

31. M J E Golay, "Hexagonal Parallel Pattern Transformations," IEEE Trans on Comp, 18(8), (1969), 733-740

32. S Gray, "Logical Properties of Binary Images in Two Dimensions," IEEE Trans on Comp 20(5), (1971), 551-561

33. H Hadwiger, Vorlesungen uber Inhalt, Oberflache und Isoperimetrie, Springer-Verlag, Berlin-Gottingen-Heidelberg, 1957

34. _____, Altes und Neues uber Konvexe Korper, Birkhauser Verlag, Basel-Stuttgart, 1955

35. H Hanckock, Development of the Minkowski Geometry of Numbers, McMillan, New York, 1939

36. D O Hebb, The Organization of Behavior, Wiley and Sons, New York, 1949

37. M Henle, A Combinatorial Introduction to Topology, Freeman, San Fransicso, 1979

38. M B Herscher and T P Kelley, "Functional Electronic Model of the Frogs Retina," IEEE Trans on Military Electronics, 7, (1963), 98-102

39. W Holsztynski, "An Introduction to the Theory of Translation Invariant Transformations and Cytocomputers," Preprint of an ERIM Tech Report, 1977

40. D H Hubel, "The Visual Cortex of the Brain," Scientific American, 5, 1963

41. M Ingram and K Preston, "Automatic Analysis of Blood Cells," Scientific American, Nov 1970

42. C V Kameswara, B Prasada, and K R Sarma, "A Parallel Shrinking Algorithm for Binary Patterns," Comp Graphics Image Processing, 5 (1976), 265-270

43. W H Kautz, "Testing for Faults in Combinational Cellular Logic Arrays," Proc 8th Annual Symp on Switching and Automata Theory, Austin TX, 1967

44. M G Kendall and P A Moran, Geometrical Probabilities, Hafner, 1963

45. D G Kendall and E F Harding, Ed, Stochastic Geometry, Wiley, New York, 1974

46. S C Kleene, "Representation of Events in Nerve Nets and Finite Automata," Automata Studies, Princeton Univ Press, Princeton, 1956

47. J Klein and J Serra, "The Texture Analyser," J of Microscopy, 95(2), 1972, 349-356

48. S R Kosaraju, "On Open Problems in the Theory of Cellular Automata," IEEE Trans on Comp, C-23, (1974), 561-565

49. B Kruse, "A Parallel Picture Processing Machine," IEEE Trans Comput, C-22, (1973), 1075-1085

50. C Y Lee, "Synthesis of a Cellular Universal Machine Using the 29-Stak Model of non-Neumann," Automata Theory Notes, Univ of Michigan Engin Summer Conf, 1964

51. E Lieblein, "Synthesis of Monogenic Iterative Arrays of Machines," ECOM-3429 Res and Dev Tech Rep, US Army, 1971

52. R M Longheed and D McCubrey, "The Cytocomputer: A Practical Pipelined Image Processor," Proc 7th Int Symp on Comp Architecture, IEEE, 1980

53. L A Lusternik, Convex Figures and Polyhedra, Heath and Co, Boston, 1966

54. A Maruoka, "Cellular Automata," J Inst Electr Commun Eng Japan, 61 (10), 1978

55. G Matheron, "Random Sets Theory and Its Applications to Sterology," J of Microscopy, 95(1), (1972), 15-23

56. G Matheron and J Serra, "Morphologic Mathematique et Granulometries on Place," Annas des Mines, 11, (1967), 39-58

57. G Matheron, Random Sets and Integral Geometry, Wiley and Sons, New York, 1976

58. W S McColloch, W H Pitts, et al, "What the Frog's Eye Tells the Frog's Brain," Proc IRE, (1959), 1940-1951

59. F Meyer, Cytologic Quantitative et Morphologic Mathematique, PhD Thesis, L'Ecole Nationale Supervieuve des Mines, Paris, 1979

60. H Minkowski, Gesammelte Abhandlungen, Teubner, Leipzig-Berlin, 1911

61. _____, Geometrie der Zahlen, Chelsea Publ, New York, 1953

62. R C Minnick and R A Short, "Cellular Linear-Input Logic," Final Report, Contract AF19(628)-498, SRI Proj 4122, Stanford Res Inst, 1964

63. R C Minnick, et al, "Cellular Arrays for Logic and Storage," Final Report, Contract AF19(628)-4233, SRI Proj 5087, Stanford Res Inst, 1966

64. R C Minnick, "Survey of Microcellular Research," Scientific Rep 1, Contract AF19(628)-5828, SRI Proj 5876, Stanford Res Inst, 1966

65. M L Minsky and S Papert, Perceptrons: An Introduction to Computational Geometry, MIT Press, Cambridge, MA, 1969

66. E F Moore, "Machine Models of Self-Reproduction," Proc Symp Appl Math, 14, AMS, (1962), 17-34

67. J C Murtha, "Highly Parallel Information Processing Systems," Advances in Computers, 7, Academic Press, New York, 1966

68. J P Mylopoulos and T Pavlidis, "On the Topological Properties of Quantized Spaces," J Assoc Comp Mach, 18, (1971), 239-254

69. A Nerode, "Linear Automation Transformations," Proc Amer Math Soc, 9, (1958), 541-544

70. M H A Newman, Elements of the Topology of Plane Sets of Points, Cambridge University Press, Cambridge, 1961

71. H Nishio, "A Classified Bibliography on Cellular Automata Theory--With Focus on Recent Japanese References," Proc IEEE Int Symp Uniformly Structured Automata and Logic, (1975), 205-214

72. S L Polyak, The Retina, University of Chicago Press, 1941

73. K Preston, "Feature Extraction by Golay Hexagonal Pattern Transforms," IEEE Trans on Comp, 20(9), (1971), 1007-1014

74. D Preston and N Norgren, "Interactive Image Processor Speeds Pattern Recognition by Computer," Electronics, 23, (1972)

75. K Preston, et al, "Basics of Cellular Logic with Some Applications in Medical Image Processing," Proc IEEE, 65(5), (1979), 820-856

76. K Preston, "Some Notes on Cellular Logic Operation," IEEE Trans on Pattern Analysis and Mach Intell, 3(4), (1981), 476-481

77. R Ransom, "Cell Interactions During Development: Array Bound Simulation of Cell Clone Growth," Progress in Cybernetics and Systems Research, 3, Wiley and Sons, New York, 1978

78. A Rosenfeld and A C Kak, Digital Image Processing, Academic Press, New York-London, 1976

79. A Rosenfeld, "Digital Topology," MAA Amer Math Monthly, 86,(1979), 621-630

80. A Rosenfeld and C R Dyer, "Cellular Pyramids for Image Analysis," T R 544 and 569, Dept Comp Sc, Univ of Maryland

81. W Rudin, Real and Complex Analysis, McGraw-Hill, New York, 1966

82. J Serra, Lectures on Image Analysis by Mathematical Morphology, Fontainbleuu France, 1976

83. A R Smith, "Cellular Automata and Formal Languages," Conf Record IEEE 11th Annual Symp on Switching and Automata Theory, (1970), 216-224

84. A R Smith, "Two-dimensional Formal Languages and Pattern Recognition by Cellular Automata," Confer Record IEEE 12th Annual Symp on Switching and Automata Theory, (1971), 144-152

85. S Sternberg, "Cytocomputer Real-time Pattern Recognition," Proc 8th Automatic Imagery Pattern Recognition Symp, (1978), 205

86. S R Sternberg, "Parallel Architectures for Image Processing," Proc 3rd Intern IEEE Compsac, 1979

87. S Sternberg, "Cytocomputer Real-time Pattern Recognition," Proc 8th Automatic Imagery Pattern Recognition Symp, (1978), 205

88. S Sternberg, "Language and Architecture for Parallel Image Processing," Proc Conf on Pattern Recognition in Practice, Amsterdam, North Holland Publ, 1980

89. S Sternberg, "Cellular Computers and Biomedical Image Processing," Proc US-France Seminar on Biomedical Image Processing, Grenoble, France, 1980

90. S L Tanimoto, Ed. Structured Computer Vision, Academic Press, New York, 1980

91. J W Thatcher, "Universality in the von Neumann Cellular Model," TR 03105-30-T, ORA, Univ of Michigan, 1964

92. J Tremblay and R Monohar, Discrete Mathematical Structures with Applications to Computer Science, McGraw Hill, 1975

93. S H Unger, "A Computer Oriented Toward Spatial Problems," Proc of the IRE, 46(10), (1958), 1744-1750

94. S H Unger, "Pattern Detection and Recognition," Proc of the IRE, (1959), 1737-1752

95. J von Neumann, Theory of Self Reproducing Machines, University of Illinois Press, Urbana, 1966

96. H Wang, "A Variant to Turing's Theory of Computing Machines,: J Assoc Comp Mach, 4, (1957), 63-92

97. H Yamada and S Amoroso, "A Completeness Problem for Pattern Generation in Tessellation Automata," J Comp Systems Sc, 4, (1970), 137-176

98. M Yoeli, "The Cascade-Parallel Decomposition of Sequential Machines,' IRE Trans EC, 10, (1963), 322-324

99. _____, "Generalized Cascade Decompositions of Automata," J ACM 12, (1965), 411-422

100. E C Zeeman, "The Topology of the Brain and Visual Perception," Topology of 3-Manifolds, Prentice Hall, Englewood Cliffs, 1962

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

TURBULENT BOUNDARY LAYERS OVER ROUGH SURFACES IN HYPERSONIC FLOW

Prepared by:	Dr. John M. Russell
Academic Rank:	Assistant Professor
Department and University:	School of Aerospace, Mechanical and Nuclear Engineering, University of Oklahoma
Research Location:	Air Force Wright Aeronautical Laboratories, Flight Dynamics Division, High Speed Aero Performance Branch
USAF Research Colleague:	Dr. Anthony W. Fiore
Date:	August 13, 1982
Contract No:	F49620-82-C-0035

TURBULENT BOUNDARY LAYERS OVER ROUGH SURFACES IN HYPERSONIC FLOW

by

John M. Russell

ABSTRACT

A practical method based on earlier contributions by Van Driest for predicting the downstream development of the skin friction, heat transfer, and momentum thickness is derived and discussed. The most obvious method for incorporating roughness effects based on a straightforward generalization of a well-known method developed in low speed flows is criticized for its lack of consideration of radiated Mach waves. A linear theory for the supersonic flow of an inviscid shear layer over a wavy wall is presented and discussed as an idealized model for steady Mach wave radiation (and its effect on skin friction) from stationary obstacles.

Acknowledgments

The author is indebted to the Air Force Office of Scientific Research and the Air Force Systems Command for sponsoring the work reported herein. He is grateful to the personnel of the High Speed Aero Performance Branch, Flight Dynamics Laboratory, Wright Aeronautical Laboratories at Wright-Patterson AFB for supplying him with a pleasant working environment, access to computer and secretarial services and moral support. The author is especially indebted to Dr. Anthony W. Fiore for suggesting the problem treated in this report and for providing able guidance during its critical stages.

I. INTRODUCTION:

The flow of turbulent boundary layers over rough walls at high speeds is of considerable interest to the Air Force. The estimation and control of flow separation, skin friction, aerodynamic heating, heat transfer and other technologically important phenomena are all limited by the depth of our understanding of compressible turbulent boundary layers. The High Speed Aero Performance Branch at the Air Force Wright Aeronautical Laboratories, spurred by recent Air Force interest in a supersonic cruise missile, has initiated a long range basic research program on the mechanics of turbulence over rough surfaces in a Mach 6 flow under the direction of Dr. Anthony W. Fiore.

Though nearly all of the in-house effort on the above project is directed toward the acquisition of quality experimental data, the investigators and sponsors felt that collateral analytical modeling of the phenomena they expect to measure would be desirable for a number of reasons, some of which will be described below in the section labeled "Objectives".

Thus, when the opportunity arose to select a number of Faculty Summer Research Fellows and to assign them to priority tasks in progress in the Laboratory, the administrators elected to choose one with a background in the analytical modeling of turbulent flows to work with Dr. Fiore on the rough wall supersonic turbulent boundary layer project.

The present author, Dr. John Russell of the School of Aerospace, Mechanical, and Nuclear Engineering at the University of Oklahoma, is a graduate of the Department of Aeronautics and Astronautics at MIT with a specialy in the analytical modeling of turbulent bursts near smooth walls, unsteady three-dimensional flows with vorticity, nonlinear effects, hydrodynamic stability and applied mathematics. Though his prior research experience in compressible turbulent boundary layers was limited at best, he apparently represented the best match between the needs of the Laboratory and the pool of skills available among this year's applicants for Faculty Summer Research Fellowships.

II. Objectives

The present ten-week research effort involved both immediate and

longer range objectives. The approaches taken in response to the immediate objectives are distinct from the approaches taken in response to the longer range ones so we will describe each set of objectives separately.

The immediate objective was to obtain a numerical prediction of the outcome of a set of experiments in progress on a milled roughness flat plate Mach 6 turbulent boundary layer at AFWAL. The planned experiments will measure the downstream development of skin friction, heat transfer, integral boundary layer thicknesses in addition to the mean velocity and density profiles. Some laser doppler and hot wire measurements of fluctuating quantities are also planned. Numerical prediction of all these quantities is desired. Disagreement between prediction and experiment is useful in isolating weaknesses of current prediction methods, while agreement reinforces the credibility of both the theory and the experiment. The possibility of discovering errors in the experiments by such comparisons must not be ruled out, either.

The longer range objective is to gain a deeper understanding of the specific physical processes that control the main qualitative features of the flow and to illuminate the manner by which two or more nominally distinct physical processes may interact. The subject of compressible turbulent boundary layers is reviewed by Bradshaw.[1] We may list four representative problem areas in which the present author's knowledge is far from complete and in which the literature he has consulted has offered less than definitive insight:

(i) How to quantitatively estimate acoustic energy radiation in the mean flow energy balance in the hypersonic range $M_\infty > 5$.

(ii) How to properly model the effects of roughness on the mean velocity and temperature profile for arbitrary Mach number and heat transfer.

(iii) What is the role of traveling disturbances (either eddies or waves) in turbulence generation and maintenance.

(iv) What is the order of importance in which various "hypersonic effects" (i.e., those that are usually neglected for $M_\infty < 5$) gain significance as M_∞ increases beyond five.

As an example of problem number (ii), we might ask whether acoustic radiation becomes important sooner or later than low Reynolds number effects associated with a sublayer that is thick compared to the overall boundary layer thickness.

We did not expect to settle all these issues in any one-man ten-week research project. Nevertheless, we directed our efforts toward the above problem areas since they seemed to offer the greatest potential for progress in the long term.

III. Prediction of Skin Friction, Heat Transfer and Boundary Layer Growth

The shape of the velocity and temperature profiles in a compressible turbulent boundary layer can be predicted with engineering accuracy by a well-known method associated with the name Van Driest. A lucid derivation of this method that relies on fewer controversial assumptions than Van Driest's original work (such as the unnecessary use of Prandtl's mixing-length concept) may be found in two papers by Rotta.[2,3] We will not reproduce the details of Rotta's derivation here. The key assumptions used in deriving the working formulas are easy to state, however, so a heuristic derivation of the main working formulas will be given in this section.

Rotta restricts attention to the portion of the boundary layer above the sublayer but deep enough within the boundary layer so that the mean streamlines are effectively parallel to the wall (or the mean y-level in the case of a rough wall). This region corresponds to the logarithmic region in low speed flows. In it, the total shear stress (consisting of a viscous part plus a Reynolds stress) is nearly independent of y and is equal to the wall shear stress τ_w. Since we are considering the region above the sublayer, the viscous stress may be neglected as well, so that

$$-\overline{\rho uv} \approx \tau_w \qquad (1)$$

(where ρ is the total instantaneous mass density of the fluid and (u,v) are the streamwise and spanwise components of the fluctuation velocity, respectively). Rotta defines a turbulent Prandtl number σ_t by the relation

$$\sigma_t \equiv \frac{\left(\frac{-\overline{\rho uv}}{\frac{dU}{dy}}\right)}{\left(\frac{-\overline{\rho vT}}{\frac{d\bar{T}}{dy}}\right)} \quad , \tag{2}$$

(where $\bar{T}$ is the mean temperature, T is the instantaneous temperature, and U is the mean streamwise velocity), and assumes that σ_t is independent of height y above the wall.

It follows from (1) and (2) that

$$-\overline{\rho vT} \simeq \frac{1}{\sigma_t} \tau_w \frac{d\bar{T}}{dU} \quad . \tag{3}$$

Now consider the energy balance for a cross sectional fluid slab illustrated by the cross-hatched region in Figure 1.

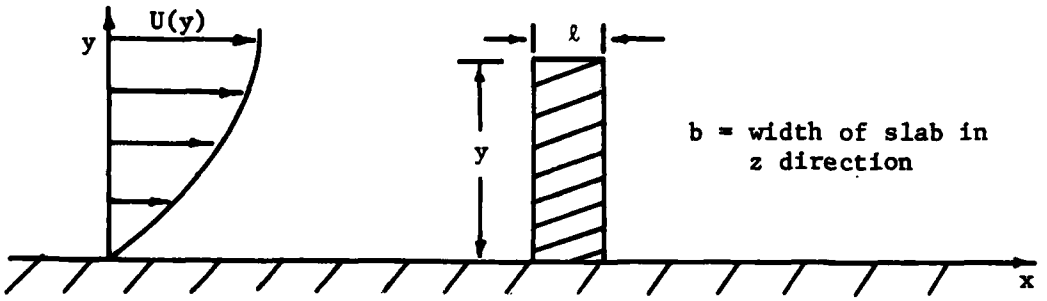


Figure 1. Cross sectional fluid slab for deriving simplified energy balance.

The rate of mechanical work $\dot{W}/\ell b$ done on the fluid slab by the surroundings (per unit of horizontal area) is dominated by the shear work, i.e.,

$$\frac{\dot{W}}{\ell b} \simeq U(y) \cdot (-\overline{\rho uv}) \simeq U(y)\tau_w \tag{4}$$

Let the rate of heat loss from the slab to the wall (per unit of horizontal area) be denoted by q_w. Finally, assuming the top of the slab is in the "fully turbulent" region, the dominant contribution to the rate of energy transfer (per unit of horizontal area) to the slab across the top of the slab is the mean flux of static enthalpy $-c_p\overline{\rho vT}$ (where c_p is the (constant) specific heat at constant pressure).

Neglecting any imbalance between the energy flux across the vertical faces of the slab, the operational form of the equation of total energy conservation becomes

$$-c_p\overline{\rho vT} = q_w - U\tau_w \tag{5}$$

Eliminating $\overline{\rho vT}$ between this equation and equation (3), we get

$$\frac{1}{\sigma_t} c_p \tau_w \frac{d\bar{T}}{dU} = q_w - U\tau_w \tag{6}$$

which is identical to equation (31) of Rotta's first paper[1] in the limit $y^+ \to \infty$ which Rotta derived by a much more thorough but laborious procedure.

Equation (6) shows that $\bar{T}$ must be a quadratic function of U. Define the parameters

$$\beta_q \equiv \frac{q_w}{\bar{\rho}_w c_p \bar{T}_w u_\tau} \tag{7a}$$

$$M_\tau = \frac{u_\tau}{a_w} \tag{7b}$$

$$u_\tau \equiv \sqrt{\frac{\tau_w}{\bar{\rho}_w}} \tag{7c}$$

$$a_w = \sqrt{\frac{\gamma \bar{p}_w}{\bar{\rho}_w}} = \sqrt{\frac{c_p}{c_v}(c_p - c_v)\bar{T}_w} = \sqrt{(\gamma-1)c_p\bar{T}_w} \tag{7d}$$

(where the subscript "w" refers to the wall level $y = 0$, c_v is the specific heat at constant volume, $\gamma \equiv c_p/c_v$, and $\bar{p}$ is the mean pressure which we will take to be a constant throughout the flow in accordance with the usual boundary layer approximation).

The integral of (6) may be written

$$\frac{\bar{T}}{\bar{T}_w} = \sigma_t \beta_q \frac{U}{u_\tau} - \sigma_t \frac{\gamma-1}{2} M_\tau^2 \left(\frac{U}{u_\tau}\right)^2 + C_1 \tag{8}$$

where $C_1 = C_1(\beta_q, M_\tau)$. Since we will have no further need to distinguish between the mean and instantaneous temperatures, we will drop the overbars on the T's in what follows.

The last key assumption which permits the derivation of a formula

relating U to y is that the velocity gradient obey the relation

$$\frac{dU}{dy} = \sqrt{\frac{\tau_w}{\bar{\rho}}}\,\frac{1}{\kappa y} \tag{9}$$

where κ is the Kármán constant. The only difference between (9) and its incompressible couterpart is that $\bar{\rho}$ is now a function of y rather than a constant. From constancy of the pressure and the equation of state of an ideal gas, we have

$$\bar{\rho}T = \bar{\rho}_w T_w \quad . \tag{10}$$

The mean density is therefore expressible in terms of the mean temperature and vice-versa.

Eliminating T between (8) and (10), one obtains an expression for $\bar{\rho}$ in terms of U. Equation (9) is then a "variables separable" ordinary differential equation between U and y:

$$\frac{dy}{\kappa y} = \frac{d(\frac{U}{u_\tau})}{\sqrt{\sigma_t \beta_q \frac{U}{u_\tau} - \sigma_t \frac{\gamma-1}{2} M_\tau^2 (\frac{U}{u_\tau})^2 + C_1}}$$

$$= \frac{d(\frac{U}{u_\tau})}{\sqrt{\sigma_t \frac{\gamma-1}{2} M_\tau^2}\ \sqrt{\frac{2C_1}{(\gamma-1)M_\tau^2\sigma_t} + (\frac{\beta_q}{(\gamma-1)M_\tau^2})^2 - (\frac{U}{u_\tau} - \frac{\beta_q}{(\gamma-1)M_\tau^2})^2}} \tag{11}$$

The substitution

$$\frac{U}{u_\tau} - \frac{\beta_q}{(\gamma-1)M_\tau^2} = \sqrt{\frac{2C_1}{(\gamma-1)M_\tau^2\sigma_t} + (\frac{\beta_v}{(\gamma-1)M_\tau^2})^2}\ \sin\theta \tag{12}$$

reduces (11) to

$$\frac{dy}{\kappa y} = \frac{1}{\sqrt{\sigma_t \frac{\gamma-1}{2}} M_\tau}\, d\theta \tag{13}$$

which has the integral

$$\theta - \theta_o = \sqrt{\sigma_t \frac{\gamma-1}{2}} M_\tau\ (\frac{1}{u} \ln \frac{y u_\tau}{\nu_w} + C_2) \tag{14}$$

in which one of the constants may be chosen arbitrarily with no loss of generality. Note that $C_2 = C_2(M_\tau, \beta_q)$. It is convenient to choose θ_o such that $U = 0$ at $\theta = \theta_o$. With this choice, it follows from (12) that

$$\sin \theta_o = \frac{-\sqrt{\frac{\sigma_t}{2(\gamma-1)}} \frac{\beta_q}{M_\tau}}{\sqrt{C_1 + \frac{\sigma_t}{2} \frac{\beta_q^2}{(\gamma-1)M_\tau}}} \tag{15}$$

$$\rightarrow \cos \theta_o = \frac{\sqrt{C_1}}{\sqrt{C_1 + \frac{\sigma_t}{2} \frac{\beta_q^2}{(\gamma-1)M_\tau}}} \tag{16}$$

Noting that $\sin \theta = \sin(\theta - \theta_o + \theta_o) = \sin(\theta - \theta_o) \cos \theta_o + \cos (\theta - \theta_o) \sin \theta_o$, equations (12)-(16) give, finally,

$$\frac{U}{u_\tau} = \sqrt{\frac{2}{(\gamma-1)M_\tau^2 \sigma_t}} \{\sin[\sqrt{\sigma_t} \frac{\gamma-1}{2} M_\tau (\frac{1}{u} \ln \frac{y u_\tau}{\nu_w} + C_2)] \sqrt{C_1}$$

$$+ (1 - \cos[\sqrt{\sigma_t} \frac{\gamma-1}{2} M_\tau (\frac{1}{u} \ln \frac{y u_\tau}{\nu_w} + C_2)] \frac{\sigma_t}{2(\gamma-1)} \frac{\beta_q}{M_\tau}\} \tag{17}$$

Equation (17) is the generalization of the "law of the wall" to the compressible heat conducting case.

An equivalent frequently quoted form may be obtained by expressing θ and θ_o in equation (14) by means of equations (12) and (15), respectively and by introducing the "incompressible" velocity u^*/u_τ by

$$\frac{u^*}{u_\tau} = \frac{1}{u} \ln \frac{y u_\tau}{\nu_w} + C_2 \qquad \text{(inner layer)} \tag{18}$$

Equation (14) then becomes

$$\frac{u^*}{u_\tau} = \frac{1}{\sqrt{\sigma_t \frac{\gamma-1}{2}} M_\tau} [\sin^{-1}(\frac{\sqrt{\frac{\sigma_t}{2}}(\gamma-1)M_\tau (\frac{U}{u_\tau} - \frac{\beta_q}{(\gamma-1)M_\tau^2})}{\sqrt{C_1 + \frac{\sigma_t}{2} \frac{\beta_q}{(\gamma-1)M_\tau^2}}})$$

$$+ \sin^{-1}\left(\frac{\sqrt{\frac{\sigma_t}{2}(\gamma-1)}\, M_\tau \frac{\beta_q}{(\gamma-1)M_\tau^2}}{\sqrt{C_1 + \frac{\sigma_t}{2}\frac{\beta_q^2}{(\gamma-1)M_\tau^2}}}\right)] \qquad (19)$$

which is the "Van Driest transformation".

The above equations are valid only within the "inner" portion of the turbulent boundary layer, say $y/\delta < .2$ where δ is the overall boundary layer thickness. Coles[4] has shown that in low speed flows, the velocity distribution for nearly all turbulent boundary layers can be fitted to an equation of the form

$$\frac{u^*}{u_\tau} = \frac{1}{\kappa}\ln\frac{yu_\tau}{\nu_w} + C_2 + \frac{\Pi}{\kappa}\left(1 - \cos\frac{\pi y}{\delta}\right) \qquad \left(\begin{matrix}\text{whole}\\ \text{layer}\end{matrix}\right) \qquad (20)$$

by suitable choice of the constants Π, δ, u_τ and where $(\kappa, C_2) = (.41, 5.0)$. Bradshaw[1] therefore suggests that equation (17) be generalized to

$$\frac{U}{u_\tau} = \sqrt{\frac{2}{(\gamma-1)M_\tau^2\sigma_t}}\left\{\sin\left(\sqrt{\sigma_t\frac{\gamma-1}{2}}\, M_\tau \frac{u^*}{u_\tau}\right)\sqrt{C_1}\right.$$

$$\left. + \left(1 - \cos\left(\sqrt{\sigma_t\frac{\gamma-1}{2}}\, M_\tau \frac{u^*}{u_\tau}\right)\right)\frac{\sigma_t\,\beta_q}{2(\gamma-1)\,M_\tau}\right\} \qquad (21)$$

with u^*/u_τ given by equation (20), so that $\frac{U}{u_\tau}$ is now specified for all values of y between the sublayer and the outer edge, provided suitable empirical functions $C_1(M_\tau, \beta_q)$ and $C_2(M_\tau, \beta_q)$ can be found. Bradshaw[2] suggests

$$C_1 = 1 \quad , \quad C_2 = 5.2 + 95\, M_\tau^2 + 30.7\, \beta_q + 226\, \beta_q^2 \qquad (22a,b)$$

The above equations may be used to derive skin friction formulas which, in turn, form the basis of a calculation method for calculating the downstream growth of the momentum thickness θ defined by*

$$\theta \equiv \frac{1}{\bar{\rho}_\delta U_\delta^2}\int_0^\delta \bar{\rho}U(U_\delta - U)\,dy \qquad (23)$$

Note that the "friction velocity Mach number" M_τ may be written in the

*This θ should not be confused with the θ introduced in equation (12).

form

$$M_\tau = \frac{u_\tau}{a_w} = \sqrt{\frac{\tau_w}{\rho_w}}\sqrt{\frac{\bar{\rho}_w}{\gamma p_w}} = \sqrt{\frac{\tau_w}{\bar{\rho}_\delta}\frac{\bar{\rho}_\delta}{\gamma \bar{p}_\delta}} = \sqrt{\frac{U_\delta^2}{2}\, c_f\, \frac{1}{a_\delta^2}} \quad ,$$

or

$$M_\tau = \sqrt{\frac{c_f}{2}}\, M_\delta \quad , \tag{24}$$

where the subscript "δ" denotes conditions at the outer edge of the boundary layer, c_f is the usual skin friction coefficient defined by

$$\tau_w = \frac{\bar{\rho}_\delta U_\delta}{2}\, c_f \quad , \tag{25}$$

and where we have used the assumption $\bar{p}_w = \bar{p}_\delta$.

In general, M_δ will be given quantity. From (24), we see that M_τ and c_f contain the same basic information. In view of the ubiquity of the parameter M_τ in the equations already derived, we will use it instead of c_f in deriving "skin friction" equations.

The quantity $\bar{\rho}/\bar{\rho}_\delta$ in (23) may be expressed in terms of U by means of equations (10) and (8):

$$\frac{\bar{\rho}}{\bar{\rho}_\delta} = \frac{\bar{\rho}_w}{\bar{\rho}_\delta}\frac{\bar{\rho}}{\bar{\rho}_w} = \frac{T_\delta}{T_w}\frac{T_w}{T} = \frac{T_\delta/T_w}{\sigma_t \beta_q \frac{U}{u_\tau} - \sigma_t \frac{\gamma-1}{2} M_\tau^2 (\frac{U}{u_\tau})^2 + C_1} \tag{26}$$

Note also that

$$\frac{u_\tau}{U_\delta} = \frac{M_\delta a_w}{M_\delta a_\delta} = \frac{M_\tau}{M_\delta}\sqrt{\frac{T_w}{T_\delta}} \quad . \tag{27}$$

It follows from (23), (26), and (27) that

$$\frac{\theta}{\delta} = \frac{M_\tau^2}{M_\delta^2}\int_0^1 \frac{\frac{U}{u_\tau}(\frac{M_\delta}{M_\tau}\sqrt{\frac{T_\delta}{T_w}} - \frac{U}{u_\tau})}{\sigma_t \beta_q \frac{U}{u_\tau} - \sigma_t \frac{\gamma-1}{2} M_\tau^2 (\frac{U}{u_\tau})^2 + C_1}\, d(\frac{y}{\delta}) \tag{28}$$

This integral may be carried out numerically using (20) and (21) provided the argument yu_τ/ν_w in (20) of the logarithm is written

$$\frac{yu_\tau}{\nu_w} = \frac{y}{\delta}\frac{\delta}{\theta}\frac{U_\delta \theta}{\nu_w}\frac{M_\tau}{M_\delta}\sqrt{\frac{T_w}{T_\delta}} \tag{29}$$

The integral (28), when evaluated, yields a relation of the form

$$\frac{\theta}{\delta} = f_1(M_\tau,\ \beta_q,\ \frac{\theta}{\delta},\ R_\theta \equiv \frac{U_\delta \theta}{\nu_w}) \tag{30}$$

provided that the parameters T_w/T_δ, M_δ, Π, κ, C_1, C_2, σ_t, and γ are specified.

Evaluating (21) at the outer edge $y/\delta = 1$, and using (27) gives

$$\frac{M_\delta}{M_\tau}\sqrt{\frac{T_\delta}{T_w}} = \sqrt{\frac{2}{(\gamma-1)M_\tau^2\sigma_t}}\left\{\sin\left(\sqrt{\sigma_t\frac{\gamma-1}{2}}\,M_\tau\,\frac{u_\delta^*}{u_\tau}\right)\sqrt{C_1}\right.$$

$$\left. + \left(1 - \cos\left(\sqrt{\sigma_t\frac{\gamma-1}{2}}\,M_\tau\frac{u_\delta^*}{u_\tau}\right)\right)\frac{\sigma_t\,\beta_q}{2(\gamma-1)\,M_\tau}\right\} \tag{31}$$

where, from (20) and (29),

$$\frac{u_\delta^*}{u_\tau} = \frac{1}{\kappa}\ln\left(\frac{\delta}{\theta}\,R_\theta\,\frac{M_\tau}{M_\delta}\sqrt{\frac{T_w}{T_\delta}}\right) + C_2 + \frac{2\Pi}{\kappa}\quad . \tag{32}$$

Evaluating (8) at $y = \delta$, we get

$$\frac{T_\delta}{T_w} = \sigma_t\beta_q\frac{M_\delta}{M_\tau}\sqrt{\frac{T_\delta}{T_w}} - \sigma_t\frac{\gamma-1}{2}M_\delta^2\frac{T_\delta}{T_w} + C_1\quad . \tag{33}$$

Equations (30)-(33) constitute three simultaneous transcendental equations for the three unknowns θ/δ, M_τ, and β_q given the local value of R_θ and the parameters listed after equation (30). These define, in essence, the "skin friction" relations for compressible heat conducting flow. The downstream growth of θ is determined from the Von Kármán momentum integral as derived, for example, by Young.[5] It reads

$$\frac{d}{dx}(\bar{\rho}_\delta U_\delta^2\theta) = -\,\delta^* U_\delta\frac{dU_\delta}{dx} + \tau_w \tag{34}$$

where δ^* is the mass weighted displacement thickness

$$\delta^* = \frac{1}{\bar{\rho}_\delta U_\delta}\int_0^\delta \bar{\rho}(U_\delta - U)dy \tag{35}$$

For flat plate flow, of course, $dU_\delta/dx = 0$, so that

$$\frac{d\theta}{dx} = \frac{\tau_w}{\bar{\rho}_\delta U_\delta^2} = \frac{c_f}{2} = \left(\frac{M_\tau}{M_\delta}\right)^2 \tag{36}$$

From the above derived "skin friction" relations, M_τ is ultimately expressible as a function of R_θ and a number of x-independent parameters, such as M_δ, κ, Π, σ_t, γ, etc. Equation (36) thus constitutes a first order ordinary differential equation for the downstream development of θ.

IV. Roughness Effects

In low speed flows, roughness effects can be incorporated conveniently by defining a nondimensional roughness height k^+ by

$$k^+ = \frac{k u_\tau}{\nu_w}$$

where k is the typical roughness height. Equation (18) can then be rewritten

$$\frac{u^*}{u} = \frac{1}{\kappa}\ln\frac{y}{k} + \frac{1}{\kappa}\ln k^+ + C_2^*(k^+)$$

$$\equiv \frac{1}{\kappa}\ln\frac{y}{k} + C_{2r}^*(k^+) \tag{37}$$

The function $C_{2r}^*(k^+)$ has been determined empirically for a variety of roughness types in adiabatic flow at zero Mach number. Rotta[6] presents a review of these data.

The most natural way to generalize these results to the compressible case is to treat the C_2 in equation (20) as a function of k^+, M_τ, and β_q, whose value at $k^+ = 0$ is given by equation (22b) and to assume that the k^+ dependence of C_2 can be represented by a multiplicentive factor determined from low speed data. Thus, we let

$$C_2(k^+, M_\tau, \beta_q) = C_2^*(k^+)\,\frac{C_2(0, M_\tau, \beta_q)}{C_2(0,0,0)} \tag{38}$$

with

$$C_2(0, M_\tau, \beta_q) = 5.2 + 95\,M_\tau^2 + 30.7\,\beta_q + 226\,\beta_q^2$$

$$C_2(0,0,0) = 5.2$$

$$C_2^*(k^+) = C_{2r}^*(k^+) - \frac{1}{\kappa}\ln k^+$$

(in accordance with equation (37)) and where $C_{2r}{}^{*}(k^{+})$ is the empirically determined function measured in adiabatic incompressible flow for the given roughness type.

V. Wave Drag of Roughness Elements

The preceding sections have been concerned with a practical calculation method suitable for immediate use. The price paid for the ability to generate mean flow predictions was a neglect of the physics of the fluctuations. In the following section, we seek to rectify this neglect by presenting a highly idealized model of Mach wave generation by roughness elements. We will study, in particular, the inviscid shear flow over a sinusoidal wavy wall in which the wave amplitude is small. The analysis permits a calculation of the wave drag of one cycle of the wall waves by a linear theory.

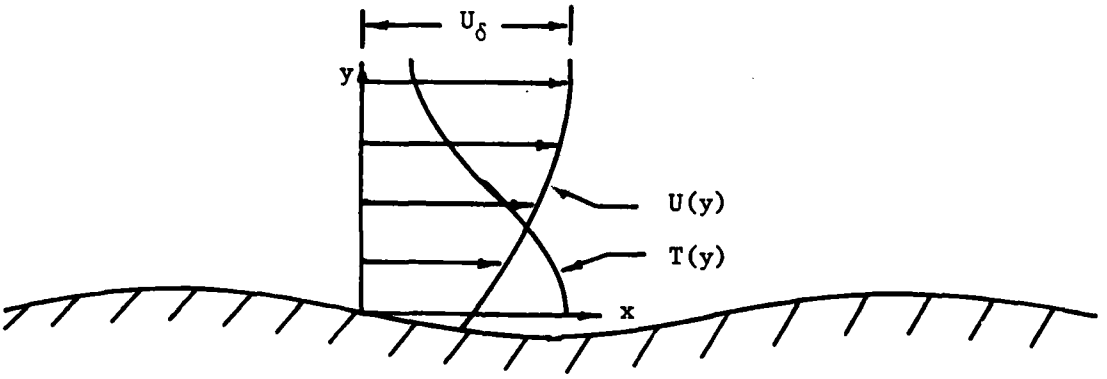


Figure 2. Shearing flow over a wavy wall: (T(y),U(y) ≡ mean temperature and velocity, respectively.

Consider an inviscid shear flow as shown in Figure 2. (x,y,z) is a right-handed coordinate system. (U(y),0,0) is the velocity field that would obtain in the absence of wall waviness. The crests of the wall waves are not necessarily parallel to the z axis. Let α and β be the streamwise and spanwise wavenumbers of the wall waves. The equation governing an isentropic change of state in an ideal gas

$$\frac{p}{p_o} = \left(\frac{\rho}{\rho_o}\right)^{\gamma}$$

is a solution of the equation

$$\frac{dp}{p} = \gamma \frac{d\rho}{\rho}$$

which applies to each fluid particle. It follows that

$$\frac{1}{p}\frac{Dp}{Dt} = \frac{\gamma}{\rho}\frac{D\rho}{Dt} \quad . \tag{39}$$

From the continuity equation*

$$\frac{D\rho}{Dt} = -\rho\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) \quad ,$$

we have

$$\frac{1}{p}\frac{Dp}{Dt} = -\gamma\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) \quad . \tag{40}$$

Combining this equation with the three components of the momentum equation

$$\frac{Du}{Dt} = -\frac{1}{\rho}\frac{\partial p}{\partial x}$$

$$\frac{Dv}{Dt} = -\frac{1}{\rho}\frac{\partial p}{\partial y} \tag{41a,b.c}$$

$$\frac{Dw}{Dt} = -\frac{1}{\rho}\frac{\partial p}{\partial z} \quad .$$

we obtain a system which when linearized for small amplitude disturbances to the basic flow yields a closed set of equation for the fluctuations. Let

$$(u,v,w,T,p,\rho) \equiv (U(y)+u',v',w',\bar{T}(y)+T',P_\infty+p',\bar{\rho}(y)+\rho') \tag{42}$$

(where T now represents the instantaneous temperature) in which P_∞ is an absolute constant. Notice that equations (40) and (41) are satisfied by the undisturbed flow with $u' = v' = w' = p' = \rho' = 0$ though the boundary condition of "no flow through the wall" is not.

It follows from the equation of state for the undisturbed flow that

$$\bar{\rho}(y)\ \bar{T}(y) = \rho_\delta\ T_\delta \tag{43}$$

---

*(u,v,w,ρ,p) now represent the local instantaneous streamwise, vertical, and spanwise velocity components, the density, and the pressure, respectively.

Introducing (42) into (40) and (41) and ignoring second-order products of the fluctuations, we obtain

$$U \frac{\partial u'}{\partial x} + \frac{dU}{dy} v' = -\frac{1}{\bar{\rho}} \frac{\partial p'}{\partial x} \tag{44}$$

$$U \frac{\partial v'}{\partial x} = -\frac{1}{\bar{\rho}} \frac{\partial p'}{\partial y} \tag{45}$$

$$U \frac{\partial w'}{\partial x} = -\frac{1}{\bar{\rho}} \frac{\partial p'}{\partial z} \tag{46}$$

$$\frac{1}{P_\infty} U \frac{\partial p'}{\partial x} = -\gamma \left(\frac{\partial u'}{\partial x} + \frac{\partial v'}{\partial y} + \frac{\partial w'}{\partial z}\right) \tag{47}$$

in which we have assumed steady flow since the wall is stationary.

Taking the x-derivative of (44) and adding the result to the z-derivative of (46), we obtain an expression involving the quantity $\partial u'/\partial x + \partial w'/\partial z$ which can be eliminated by means of (47). Thus

$$U \frac{\partial}{\partial x} \left(-\frac{1}{\gamma P_\infty} U \frac{\partial p'}{\partial x} - \frac{\partial v'}{\partial y}\right) + \frac{\partial U}{\partial y} \frac{\partial v'}{\partial x} = -\frac{1}{\bar{\rho}} \left(\frac{\partial^2 p'}{\partial x^2} + \frac{\partial^2 p'}{\partial z^2}\right) \tag{48}$$

This equation together with (45) yields a closed set for the variables v' and p'. Note that the coefficients in these equations are all independent of x and z. We therefore seek separable solutions in the form

$$v'(x,y,z) = \mathrm{Re}\{\tilde{v}(y) e^{i(\alpha x + \beta z)}\} \tag{49}$$

$$p'(x,y,z) = \mathrm{Re}\{\tilde{p}(y) e^{i(\alpha x + \beta z)}\} \tag{50}$$

in which $\tilde{v}$ and $\tilde{p}$ may be complex (even though v' and p' are real).

Substituting these expressions into (45) and (48) and demanding that the solutions be nontrivial, we obtain

$$i\alpha U \tilde{v} = -\frac{1}{\bar{\rho}} \frac{d\tilde{p}}{dy} \tag{51}$$

from (45), and

$$\left(\frac{\alpha^2 U^2}{\gamma P_\infty} - \frac{(\alpha^2 + \beta^2)}{\bar{\rho}}\right) \tilde{p} = i\alpha \left(U \frac{d\tilde{v}}{dy} - \tilde{v} \frac{dU}{dy}\right) \tag{52}$$

from (48). Eliminating $\tilde{p}$, we obtain a single equation for $\tilde{v}$

$$\frac{d}{dy} \left[\frac{U \frac{d\tilde{v}}{dy} - \tilde{v} \frac{dU}{dy}}{\frac{\bar{T}}{\bar{T}_\delta} - M_\delta^2 \frac{\alpha^2}{k^2} \left(\frac{U}{U_\delta}\right)^2}\right] = \frac{k^2}{\frac{\bar{T}}{T_\delta}} U \tilde{v} \tag{53}$$

in which we have introduced the notation $\alpha^2 + \beta^2 \equiv k^2$ and where we have written

$$\frac{\bar{\rho}_\delta}{\gamma P_\infty} \equiv \frac{1}{a_\delta^2} \quad , \quad \frac{U_\delta}{a_\delta} \equiv M_\delta \quad , \quad \text{and} \quad \frac{\bar{\rho}_\delta}{\bar{\rho}} = \frac{\bar{T}}{\bar{T}_\delta}$$

Notice that as $y \to \infty$, the above equation reduces to

$$\frac{d^2\tilde{v}}{dy^2} - k^2\left(1 - \left(\frac{\alpha}{k} M_\delta\right)^2\right)\tilde{v} = 0 \qquad (y > \delta) \tag{54}$$

which has the solutions

$$\tilde{v} \to (e^{\beta y}, e^{-\beta y}) \tag{55}$$

where

$$\beta \equiv k\sqrt{1 - \left(\frac{\alpha}{k} M_\delta\right)^2} \tag{56}$$

The solution $\tilde{v} \sim e^{-\beta y}$ corresponds to "outgoing" Mach waves and is the one of physical interest in the present problem. We therefore impose the outer edge "boundary condition

$$\frac{d\tilde{v}}{dy} + \beta\tilde{v} = 0 \qquad \text{for } y = \delta \tag{57}$$

to rule out the physically unacceptable solution proportional to $e^{\beta y}$ (which represents "incoming" Mach waves, created by some wavemaker other then the wall).

Integrating (53) with respect to y from y to δ and applying (57), we get

$$U\frac{d\tilde{v}}{dy} - \tilde{v}\frac{dU}{dy} = \left[\frac{\bar{T}}{\bar{T}_\delta} - \left(M_\delta \frac{\alpha}{k}\frac{U}{U_\delta}\right)^2\right]\left\{-\frac{-U_\delta k^2}{\beta}\tilde{v}_\delta - k^2\int_y^\delta \frac{U\tilde{v}}{\frac{\bar{T}}{\bar{T}_\delta}}\,dy'\right\} \tag{58}$$

Dividing (58) by U^2 and noting that

$$\frac{U\frac{d\tilde{v}}{dy} - \tilde{v}\frac{dU}{dy}}{U^2} = \frac{\partial}{\partial y}\left(\frac{\tilde{v}}{U}\right)$$

we may integrate again with respect to y from y to δ to obtain

$$\frac{\tilde{v}_\delta}{U_\delta} - \frac{\tilde{v}}{U} = \int_y^\delta \left[\left(\frac{\frac{\bar{T}}{\bar{T}_\delta}}{U^2} - \left(\frac{M_\delta\frac{\alpha}{k}}{U_\delta}\right)^2\right)\left\{\frac{-U_\delta k^2}{\beta}\tilde{v}_\delta - k^2\int_{y'}^\delta \frac{U\tilde{v}}{\frac{\bar{T}}{\bar{T}_\delta}}\,dy''\right\}dy'$$

or since

$$\beta = ik\sqrt{\left(\frac{\alpha}{k} M_\delta\right)^2 - 1},$$

$$\frac{\tilde{v}}{\tilde{v}_\delta} = \frac{U}{U_\delta}\left\{1+\int_{y/\delta}^{1}\left(\frac{\bar{T}}{\bar{T}_\delta}\left(\frac{U_\delta}{U}\right)^2 - \left(M_\delta\frac{\delta}{k}\right)^2\right)d\left(\frac{y'}{\delta}\right)\right. \quad \frac{-i(k\delta)}{\sqrt{\left(\frac{\alpha}{k}M_\delta\right)^2-1}}$$

$$\left. + (k\delta)^2\int_{y/\delta}^{1}\left[\left(\frac{\bar{T}}{\bar{T}_\delta}\frac{U_{\delta 2}}{U^2} - \left(M_\delta\frac{\alpha}{k}\right)^2\right)\int_{y'/\delta}^{1}\frac{\frac{U}{U_\delta}\frac{\tilde{v}}{\tilde{v}_\delta}}{\frac{\bar{T}}{\bar{T}_\delta}}d\left(\frac{y''}{\delta}\right)\right]d\left(\frac{y'}{\delta}\right)\right\} \qquad (59)$$

The exact solution to (59) can be obtained recursively by taking

$$\frac{\tilde{v}}{\tilde{v}_\delta} \simeq \frac{U}{U_\delta}\left\{1 - \frac{i(k\delta)}{\sqrt{\left(\frac{\alpha}{k}M_\delta\right)^2-1}}\int_{y/\delta}^{1}\left[\frac{\bar{T}}{\bar{T}_\delta}\frac{U_{\delta 2}}{U^2} - \left(\frac{M_\delta\alpha}{k}\right)^2\right]d\left(\frac{y'}{\delta}\right)\right. \qquad (60)$$

as the first approximation. Substituting the first approximation back into the right hand side of (59) gives a second approximation which may be further substituted to give a third approximation, and so forth. Experience in the theory of hydrodynamic stability suggests that the first approximation (60) called the "small wavenumber" approximation may be quite useful as is without any further iterations.

Note that the slope of the streamline ψ = constant satisfies the condition

$$\left(\frac{dy}{dx}\right)_{\psi=\text{const.}} = \frac{v}{U} = \text{Re}\left\{\frac{\tilde{v}(y)}{U}e^{i(\alpha x+\beta z)}\right\}$$

Let $y = 0$ be the x-averaged height of the wall. Let ψ_δ denote the stream surface whose x-averaged elevation is δ. Let $\psi = 0$ be the stream function on the wall.

Then

$$\left(\frac{dy}{dx}\right)_{\psi=\psi_\delta} = \left(\frac{v}{U}\right)_{\psi=\psi_\delta} = \text{Re}\left\{\frac{\tilde{v}_\delta}{U_\delta}e^{i(\alpha x+\beta z)}\right\}$$

$$= \left|\frac{\tilde{v}_\delta}{U_\delta}\right|\cos\left(\alpha x + \beta z + \arg\left(\frac{\tilde{v}_\delta}{U_\delta}\right)\right) \quad .$$

Therefore,

$$\max\left\{\left(\frac{dy}{dx}\right)_{\psi=\psi_\delta}\right\} = \left|\frac{\tilde{v}_\delta}{U_\delta}\right|$$

Similarly

$$\max\{(\frac{dy}{dx})_{\psi=0}\} = (\frac{\tilde{v}(0)}{U(0)}) \quad (U(0) \neq 0)$$

It follows from (60) that the ratio of the maxima of the streamline slopes is given by

$$\frac{\max\{(\frac{dy}{dx})_{\psi=\psi_\delta}\}}{\max\{(\frac{dy}{dx})_{\psi=0}\}} = \left|\frac{\tilde{v}_\delta}{\tilde{v}(0)}\frac{U(0)}{U_\delta}\right| \cdot \left|1 - \frac{i(k\delta)}{\sqrt{\alpha M_\delta{}^2 - (\frac{\alpha}{k})^2 - 1}} \int_0^1 [\frac{\bar{T}}{\bar{T}_\delta}\frac{U_\delta{}^2}{U^2} - (\frac{M_\delta \alpha}{k})^2] d(\frac{y}{\delta})\right|^{-1}, \tag{61}$$

which is calculable quantity given the distributions $U(y)$, $\bar{T}(y)$, and the parameters M_δ, α, and β. Thus, for a given wall wave, the maximum slope of the streamline at the outer edge is known.

Now in steady inviscid flow, the stream surface $\psi = \psi_\delta$ may be replaced by a rigid wall without affecting the flow above it. Thus, the wave drag exerted on the fluid below the surface $\psi = \psi_\delta$ is equal to the wave drag exerted by a nominally uniform flow $U = U_\delta$ over an equivalent wavy wall. The solution to the latter problem has been given by Liepmann and Roshko[7] in the two-dimensional case to which we will now restrict attention. Those authors show that

$$C_D \equiv \frac{\alpha}{2\pi}\int_0^{\frac{2\pi}{\alpha}} C_p (\frac{dy}{dx})_{\psi=\psi_\delta} dx = \frac{2}{\sqrt{M_\delta{}^2 - 1}} \frac{\alpha}{2\pi}\int_0^{\frac{2\pi}{\alpha}} [(\frac{dy}{dx})_{\psi=\psi_\delta}]^2 dx \tag{62}$$

where C_p is the pressure coefficient defined by

$$C_p = \frac{p - p_\infty}{\frac{\bar{\rho}_\delta}{2} U_\delta{}^2}$$

From steadiness of the flow and periodicity in the x-direction, the principle of linear momentum conservation implies that the net streamwise force per unit of horizontal area on the fluid between $\psi=\psi_\delta$ and $\psi=0$ must be zero. It follows that the drag coefficient exerted on the wall $\psi=0$ is given by equation (62).

Note that

$$\int_0^{\frac{2\pi}{\alpha}} \left[\left(\frac{dy}{dx}\right)_{\psi=\psi_\delta}\right]^2 dx = \int_0^{\frac{2\pi}{\alpha}} \left[\left(\frac{dy}{dx}\right)_{\psi=0}\right]^2 dx \cdot \left(\frac{\max\{(\frac{dy}{dx})_{\psi=\psi_\delta}\}}{\max\{(\frac{dy}{dx})_{\psi=0}\}}\right)^2 \tag{63}$$

where the quantity $(\frac{dy}{dx})_{\psi=0}$ is known from the given wall geometry and the quantity in parenthesis is calculable from equation (61).

This completes our discussion of the wave drag of a wavy wall in an inviscid supersonic shear flow.

VI. Results

It was our original intention to prepare a computer code for the calculation of the downstream development of $\theta(x)$, $c_f(x)$, and $\beta_q(x)$ and to use it to generate predictions of the outcome of the experiments in progress at AFWAL. Due to time limitations, it has not proved possible to achieve this goal during the allotted 10-week summer research period. Our "results", then, consist in the above presentation of the algorithm by means of which such calculations may be carried out in the near future.

VII. Recommendations

Based on our investigation into the calculation of turbulent boundary layers in hypersonic flow, we may make the following recommendations for future work:

1. Routine predictions of $\theta(x)$, $c_f(x)$, and $\beta_q(x)$ over rough surfaces in supersonic flow should, for the time being, be made with the aid of the algorithm presented in Sections III and IV above or one similar to it.

2. More effort should be directed toward improving the rather crude method for incorporating the roughness effects outlined in Section IV.

3. Effects of radiated Mach waves should be investigated more thoroughly. The idealized model given in Section V above ignores viscous effects and travelling wave disturbances. The idealized energy equation (5) could be significantly affected by acoustic radiation for $M_\delta > 5$, for example.

4. A fundamental study of "coherent structures" in compressible turbulent boundary layers should be initiated to investigate the qualitative effects of compressibility and heat transfer on "turbulent bursts" and other turbulence production mechanisms that have been documented in low speed flows.

REFERENCES

1. Bradshaw, P. (1977) "Compressible Turbulent Shear Layers," Annual Reviews of Fluid Mechanics, vol. 9, pp. 33-54.

2. Rotta, J.C. (1959) "Über den Einfluss der Machschen Zahl und des Wärmeübergangs auf das Wandegesetz turbulenter Strömung." Zeitschrift der Flugwissenschaften Band 7, Seite 264-274.

3. Rotta, J.C. (1960) "Turbulent Boundary Layers With Heat Transfer in Compressible Flow." A.G.A.R.D. Report 281.

4. Coles, D. and Hirst, E.A. (1968) Proceedings, 1968 AFOSR-IFP Stanford Conference on Computation of Turbulent Boundary Layers. Thermosciences Division, Stanford University.

5. Young, A.D. (1953) "Boundary Layers," chapter X of Modern Developments in Fluid Dynamics, High Speed Flow, L. Howarth, ed. Oxford University Press, p. 391.

6. Rotta, J.D. (1972) Turbulente Strömungen, B.G. Teubner Verlag, Stuttgart, section 3.2.5.

7. Liepmann, H.W. and Roshko, A. (1957) Elements of Gasdynamics John Wiley and Sons, section 8.6.

1982 USAF - SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

CONCEPTUALIZATION OF THE DYNAMIC BEHAVIOR OF THE FLOW-FIELD IN THE APL COMBUSTOR

Prepared by:	Sarwan S. Sandhu
Academic Rank:	Assistant Professor
Department and University:	Department of Chemical Engineering, University of Dayton
Research Location:	Aero Propulsion Laboratory, Wright-Patterson Air Force Base, Ohio
USAF Research Colleague:	Dr. W.M. Roquemore
Date:	July 9, 1982
Contract No.:	F49620-82-C-0035

CONCEPTUALIZATION OF THE DYNAMIC BEHAVIOR OF THE FLOW-FIELD IN THE APL COMBUSTOR

by

Sarwan S. Sandhu

ABSTRACT

Results of a brief investigation of the dynamic behavior of a bluff-body stabilized diffusion flame are presented for air and fuel flow rates of 1kg/s and 6kg/hr, respectively. A cine film, acquired at 4000 frames/s, of the reacting flow field near the bluff-body was analyzed with regard to time variant interaction between the annulus air stream and the flame stabilizing recirculation zone. Air vortices generated from the shear layer of the air stream appear to interact with the recirculating zone while growing as they move downstream. Estimated air vortex average axial and rotational velocities are 22.8 m/s and 548.9 rad/s, respectively. Estimated peripheral average angular velocity of reactive fluid recirculating zone is 506.5 rad/s. Fluid, made visible by dispersed reacting and emitting relatively "small" sized gaseous pockets, appears to move upstream in a core about the centerline towards the recirculating spatial region. Reacting fluid, made visible by light emission, appears to be peeled off the recirculating reactive zone in the form of cylindrical shells with irregular boundaries at quasi-periodic intervals. Ratio of air vortex to flame turbule frequency at an axial dimensionless location of 0.68 is about 2/1.

ACKNOWLEDGEMENTS

The author would like to thank the Air Force Office of Scientific Research and Southeastern Center for Electrical Engineering for providing him with the opportunity of a very worthwhile summer at the Aero Propulsion Laboratory, Wright-Patterson Air Force Base, Ohio. He would like to thank Dr. W.M. Roquemore for helpful discussions. Assistance of Miss Robin Archer during the course of this investigation is acknowledged. The expert typing skill of Mrs. Mary Lu Kinzeler is appreciated.

CONCEPTUALIZATION OF THE DYNAMIC BEHAVIOR OF THE FLOW-FIELD IN THE APL COMBUSTOR

I & II. Objective and Introduction

The main objective of the effort expended in the ten week summer period was to clarify the dynamic behavior of the flow-field in the APL combustor in the "near bluff-body spatial region". Such clarification helps in gaining insight into the air and fuel mixing process resulting in reactant consuming exothermic combustion process. Unsteady flame behavior was observed and studied in terms of flame turbules in the APL combustor,[1,2,3] using cine photograpic and CH emission recording techniques. Later, high speed Mn-laser shadowgraphy and cine photography were employed to further shed some light on the dynamic behavior of the APL combustor. A cine film acquired at 4000 frames per second for air flow rate of 1kg/s and fuel flow rate of 6kg/hr indicated the dynamic behavior more clearly. Dynamic behavior of the flow field was investigated by analyzing the cine and high speed shadowgraphic films.

Methods of analyses to acquire various types of qualitative and quantitative information regarding the dynamic behavior of the flow field are described. Deduced dynamic flow field is approximately depicted as contour diagrams. Quantitative estimated results on air indentation (or vortex) axial velocity, its growth rate and frequency; flame turbule frequency and length for air flow rate of 1kg/s and fuel flow rate of 6kg/hr are briefly discussed. Estimated characteristic dimensions of the CO_2 jet for air flow rate of 1kg/s and CO_2 flow rate of 4kg/hr are also reported.

III. EXPERIMENTAL

Combustion Rig

A schematic diagram of the Aero Propulsion Laboratory (APL) combustion tunnel is shown in Figure 1. The centerbody is 79 cm. long and 14 cm. in diameter. Gaseous propane fuel is injected through a 4.8 mm diameter tube at the center of the centerbody. Air flows through the annular space between the outer duct and the centerbody. A 31.8 mm long square-cell honeycomb flow straightener with a cell size of 4.8 mm square and two number 16 mesh screens are mounted in the annulus. The 25.4 cm diameter duct has 30.5 x 7.6 cm viewing ports that provide both optical and conventional probe access to combusting regions. Additional information about the combustion tunnel is given in References 1 and 3.

QUALITATIVE CHARACTERIZATION OF FLOW FIELD

The qualitative characterization of the flow field of the bluff-body diffusion flame was obtained by examining shadowgraphic and cine films of the flame at an air flow rate of 1kg/s and fuel flow rates of 4 and 6kg/hr. The shadowgraphic films were shot at 6583 frames/s and the cine films at 4000 frames/s. Two hundred photographic stills were obtained from the cine films and used for closer examination. The viewing port did not facilitate the visualization of the entire flame so that the flame was assumed to be symmetrical and only half was investigated.

QUANTITATIVE DATA COLLECTION

The cine films, shadowgraphic films, and photographic stills were used to quantify certain aspects of the flow field. In all cases, the bluff-body image diameter was measured and a magnification factor obtained to facilitate the conversion of image measurements to real values. The cine films were used to determine air indentation or vortex

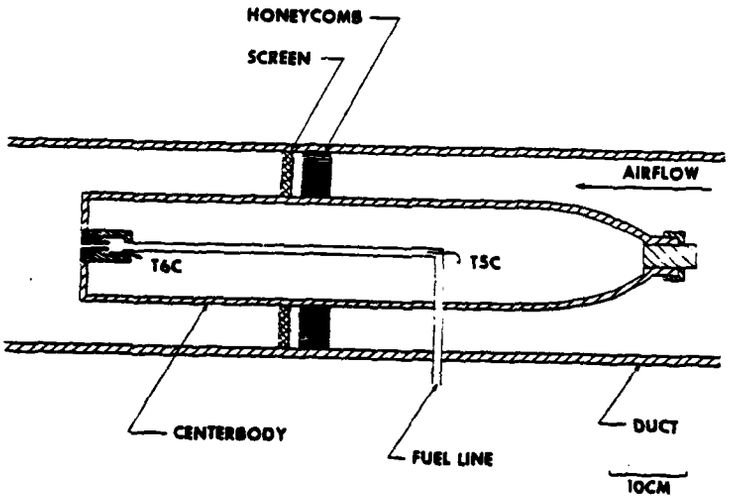


Figure 1. Combustion Rig

linear velocities, flame turbule and air vortex shedding frequencies, and recirculation zone rotational velocity. The photographic stills were used to measure flame turbule length and air vortex and recirculation zone rotational velocities. The shadowgraphic films were used to determine the cold jet dimensions for CO_2 flow rate of 4kg/hr and air flow rate of 1kg/s.

The presence of an air vortex was detected by observing the air-reaction zone interface. An indentation of the reactiong zone, caused by the vortexing air, indicated the presence of an air vortex. These indentations were used to measure air vortex shedding frequency and velocities. The presence of a flame turbule was indicated by the formation of a cylindrical shell of reacting and emitting fluid which moved downstream.

AIR VORTEX SHEDDING AND FLAME TURBULE FREQUENCIES

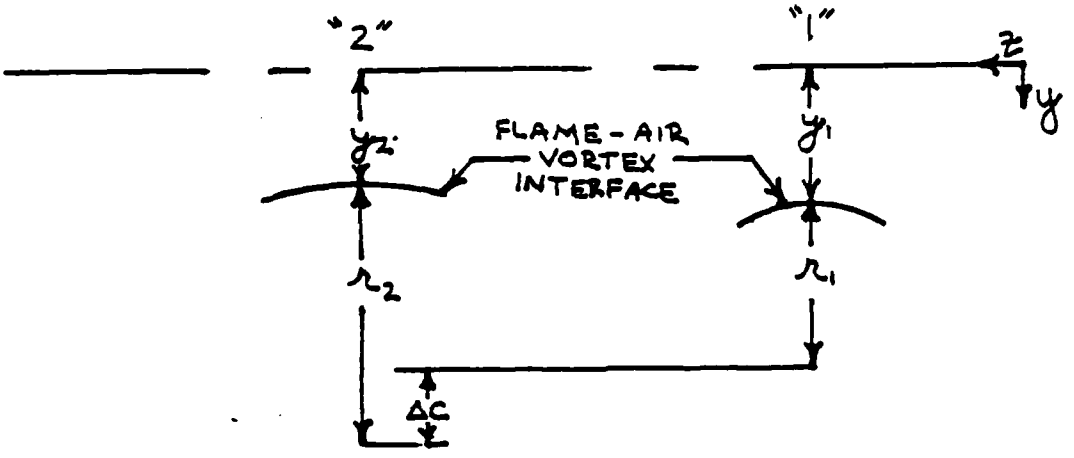
Two locations were chosen to measure air vortex shedding frequency; 5 and 9.5 cm downstream of the bluff-body face. Flame turbule frequency was determined at a location 9.5 cm downstream of the bluff-body face. The cine film was viewed at 2 frames/s to measure frequencies.

When measuring air vortex shedding frequency, the number of indentations that appeared in a set number of frames was counted; the number of indentations and number of frames were recorded. The frequency was obtained by dividing the number of indentations by the time equivalent of the number of frames. Ten and nine independent measurements were taken at the 5 and 9.5 cm locations, respectively. The arithmetic mean of the measurements was found to give an average air vortex shedding frequency at each location.

To measure flame turbule frequency, the number of reacting fluid shells which passed by the location in a set number of frames was recorded. Six independent measurements were made and those results were averaged to yield an average flame turbule frequency.

AIR VORTEX LINEAR VELOCITIES

For simplification of analyses, as well as to make the results more consistent and less subjective, each air vortex was assumed to be circular. The center and arc of the vortex were then traced through a series of frames, corresponding to a real time interval. The velocity in the Z-direction was determined by measuring the movement of the vortex center in the Z-direction. The velocity in the y-direction, the growth rate, and the flame velocity were determined in the following manner.



$$\because\ y_2 + r_2 - \Delta C = y_1 + r_1$$

$$y_2 - y_1 = \Delta C - (r_2 - r_1)$$

$$\therefore\ \frac{y_2 - y_1}{\Delta t} = \frac{\Delta C}{\Delta t} - \frac{2(r_2 - r_1)}{2\Delta t},$$

WHERE Δt IS THE TIME INTERVAL FROM "1" TO "2"

(Air vortex-flame interface radial velocity) = (air vortex center radial velocity) - 1/2 (air vortex diametrical growth rate)

The flame-air vortex interface velocity was measured by drawing horizontal tangents to the vortex curves. Air vortex center velocity in the y-direction was measured by the movement of the vortex center in the y-direction. Growth rate was determined from twice the change in radii of the vortex curves.

RECIRCULATION ZONE ROTATIONAL VELOCITY

The recirculation zone rotational velocity was determined using photographic stills under the assumption that the air vortex was an identifiable object on the recirculation zone. The angle swept by the vortex in a set number of frames was recorded. The rotational velocity was determined by dividing the angle by the time equivalent of the frames. Such measurements were performed on five distinct air vortices.

FLAME TURBULE LENGTH

The flame turbule length was found by use of the photographic stills. Appropriate frames in which the turbule was apparent were found and the length was determined directly from the picture by the application of magnification factor. Figure 2 illustrates two such appropriate frames. Four independent measurements were made and the results were averaged to give an average flame turbule length.

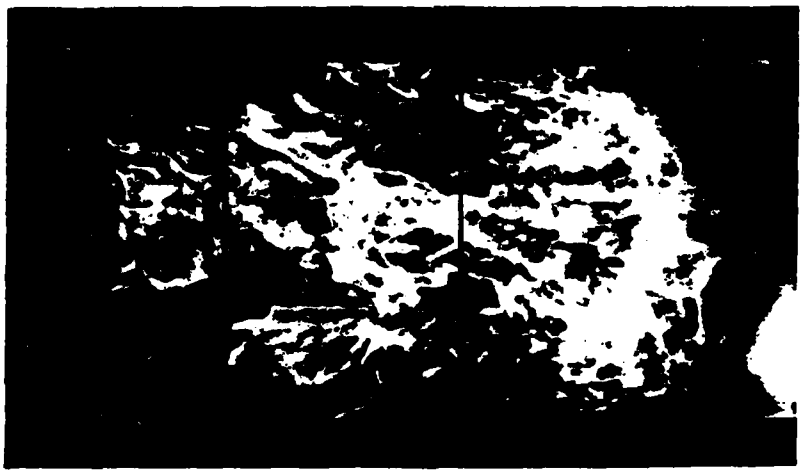
AIR VORTEX ROTATIONAL VELOCITY

The photographic stills were used to measure air vortex rotational velocity. A series of frames were selected in which an identifiable eddy was present. Figure 3 shows one such series. The angle swept from one frame to the next was measured and from this the air vortex angular velocity was determined. Eleven independent measurements were made and the arithmetic mean was calculated to give an average air vortex

Air flow rate = 1kg/s
Fuel flow rate = 6kg/hr



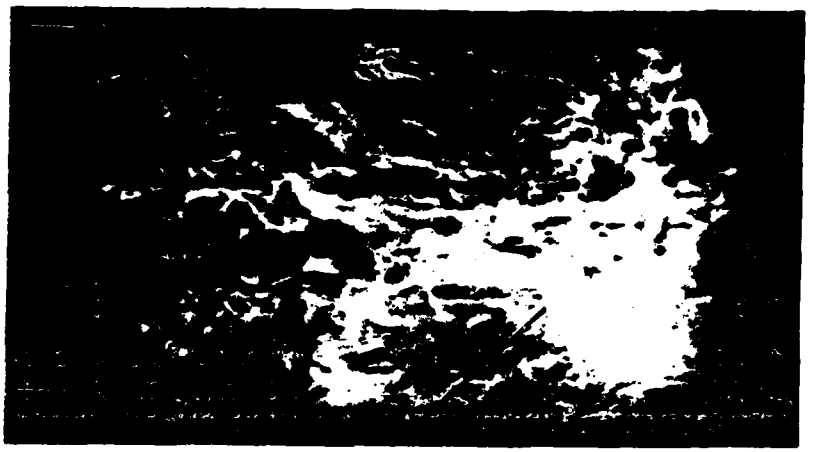
(1)



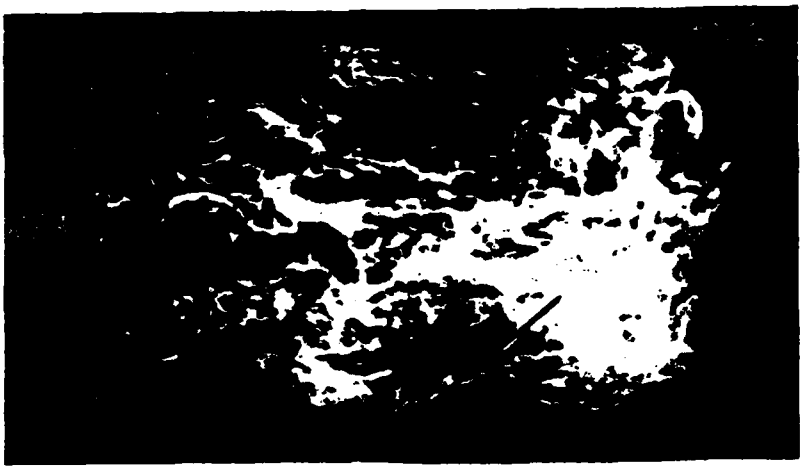
(2)

Figure 2. Photographic stills made from cine film (framing speed 4000/s) illustrating method of estimation of flame turbule length

Air flow rate = 1kg/s
Fuel flow rate = 6kg/hr

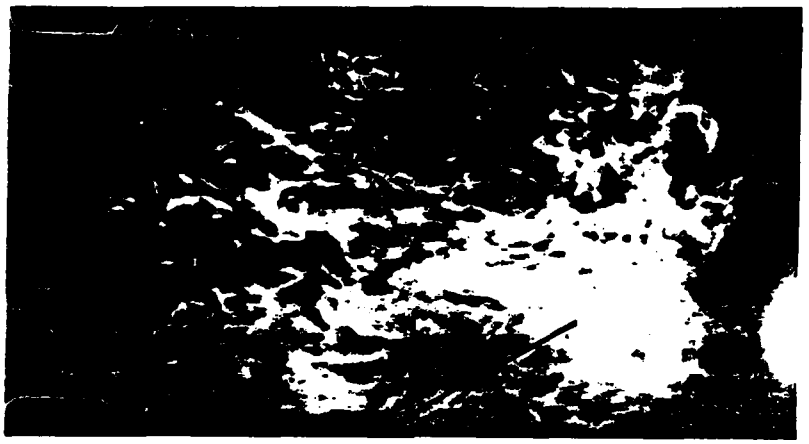


(1)

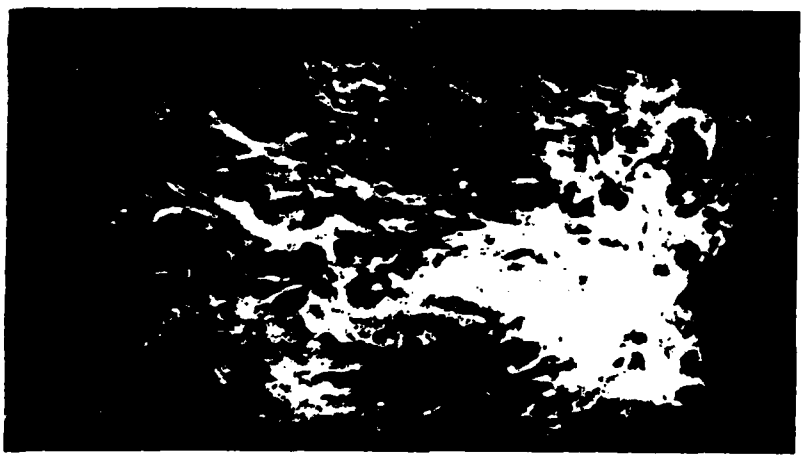


(2)

Figure 3. Photographic "stills" made from cine film (framing speed, 4000/s) illustrating method of estimation of air indentation rotational velocity.



(3)



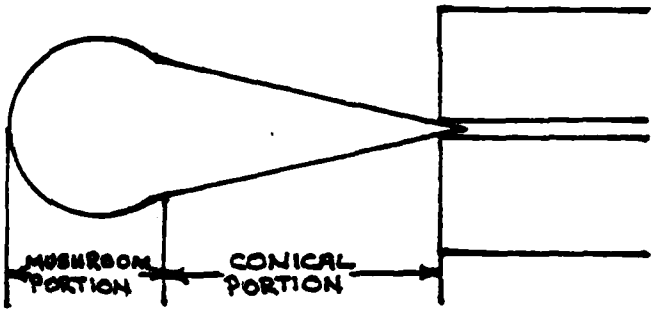
(4)

Figure 3. (continued)

rotational velocity.

COLD JET DIMENSIONS

In the shadowgrapic film, the outline of the cold jet was visible. The jet consists of a linearly expanding portion, originating at the bluff-body. This conical portion terminates in a "mushroom" portion.



The diameter at the end of the conical portion of the jet and the axial position of this diameter, the maximum jet diameter and the axial position of this diameter, and the axial position of the end of the jet were measured and recorded. Six independent measurements of each dimension were taken and averaged to yield the average cold jet dimensions.

IV. RESULTS AND DISCUSSIONS

A MODEL OF NEAR REGION REACTING FLOW FIELD IN THE WAKE OF THE BLUFF-BODY

Figure 4 depicts a model of near flow field deduced from cine and high speed shadowgraphic film analyses. A picture of the flow-field indicated is considered to be applicable for air flow rate of 1kg/s and fuel flow rate range up to 6kg/hr. Fuel jet, 1, spreads in a conical manner up to Z/D = 0.4 by entraining fluid from the surrounding environment. Fuel jet "mushroom" is apparent between Z/D = 0.4 to 0.7. Fuel jet "mushrooming" can be thought of as a resultant effect of backward flow, 2 , interaction against the forward flowing jet fuel, and inner, 3 and outer, 5 recirculation zones. Analyses of high speed shadowgraphic films indicate that fuel jet mushroom "wobbles" with time. The jet mushroom wobbling phenomenon may be explained in terms of fluid interactions and time varying turbulence structure of the jet-surrounding flow field. Jet fuel and hot backward flowing fluid, 2, are shown to mix as indicated by 4 . The hot fluid and jet fuel keep on mixing by turbulence and molecular diffusion processes while the mixed fluid moves upstream towards the face of the bluff-body, 9 . The mixed flow fluid stream, 4 splits into two portions. One becomes part of the inner recirculation zone,3, and the other part of the outer recirculation zone, 5. Structure and rotational orientation of the inner recirculation zone is indicated by 3. The jet fuel and the inner recirculation zone fluid mixing could occur via turbulent entrainment and molecular diffusion processes. Time varying structures of outer recirculation zone,5, and air indentation (or, air vortex), 6, are depicted in Fig. 4 a - e. Recirculation zone, 5 extends to almost Z/D

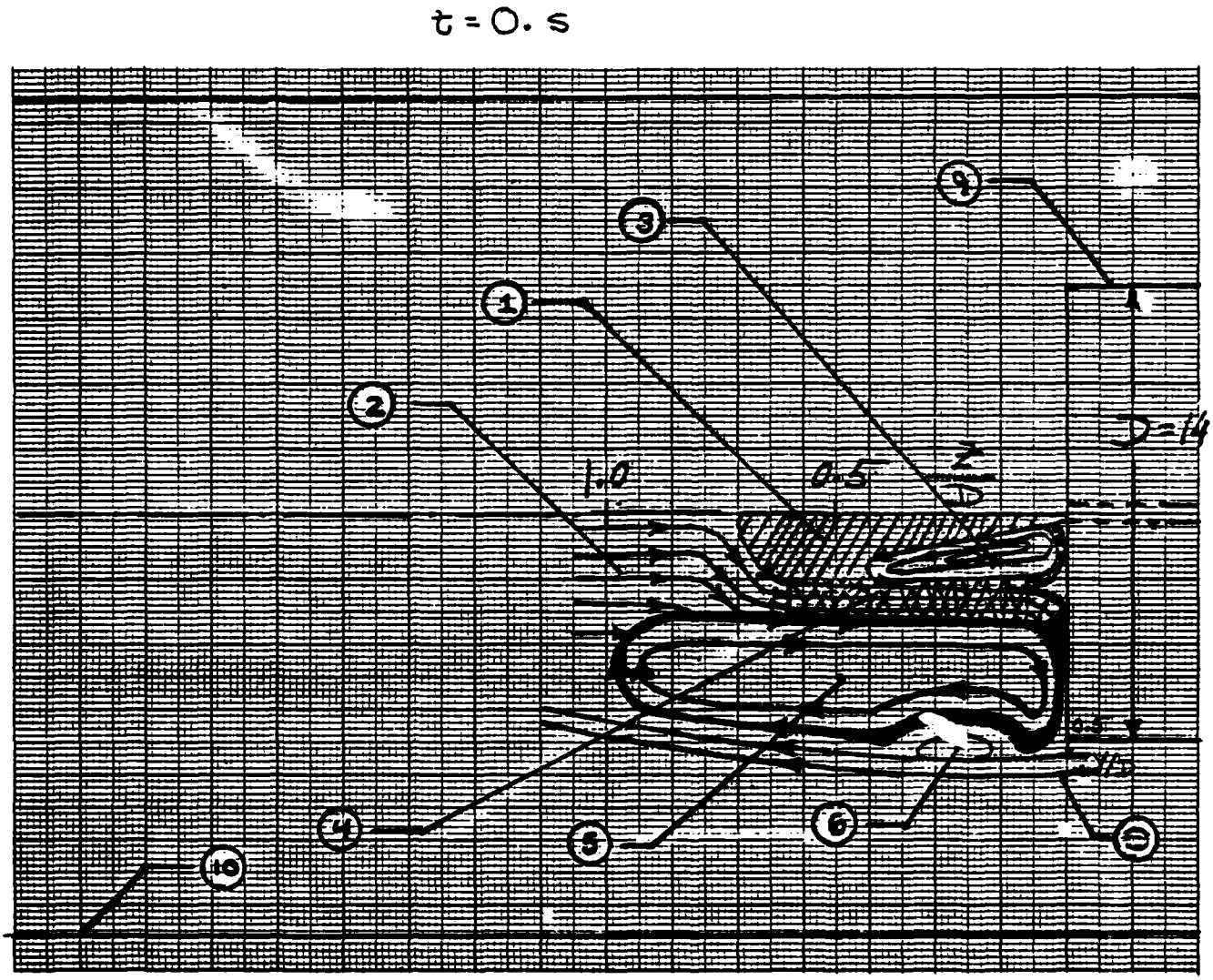


Figure 4. A conceptual model of near region reacting flow field in the wake of the bluff-body in the APL combustor. Scale: 1/2 of full size.

LEGEND

Fuel jet, 1

"Backward Flow" 2

Inner recirculation zone, 3

"Mixed Upstream Flow" 4

"Outer Recirculation Zone" 5

Air Vortex, 6

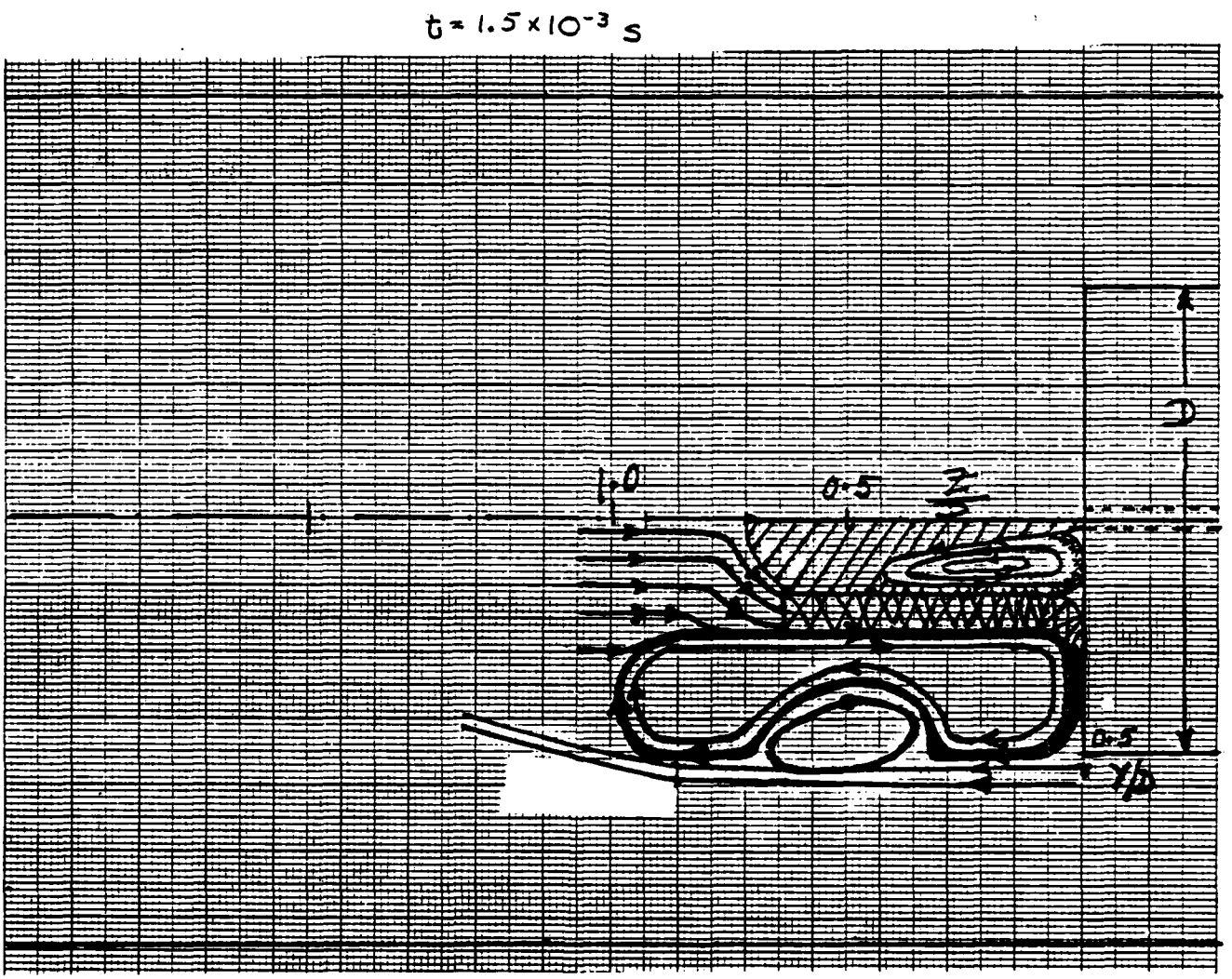
Flame turbule fragment, 7

Annulus air stream, 8

Cylindrical bluff-body incorporating fuel injection tube, 9

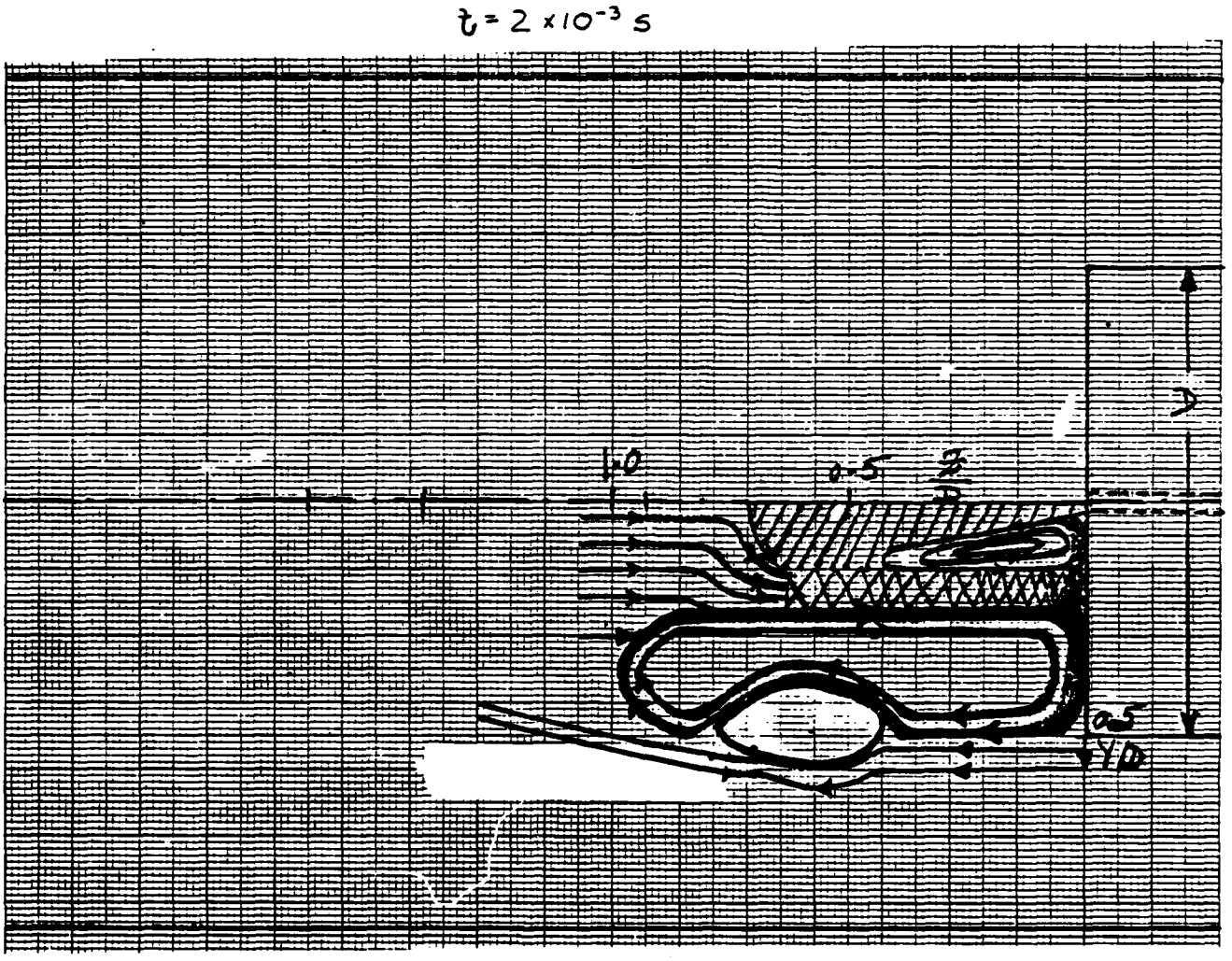
Combustion duct, 10

Figure 4 (continued)



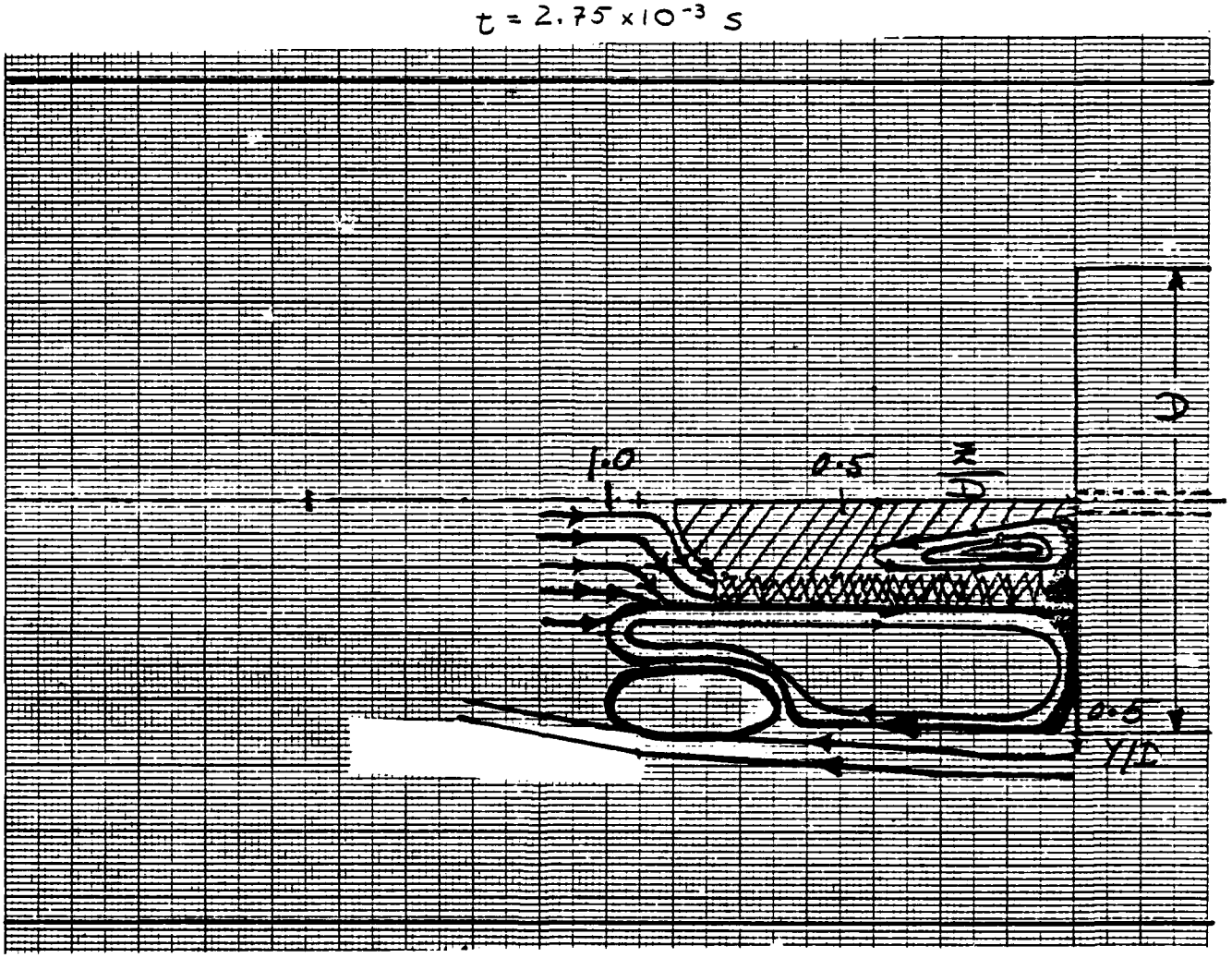
(b)

Figure 4. (continued)



(c)

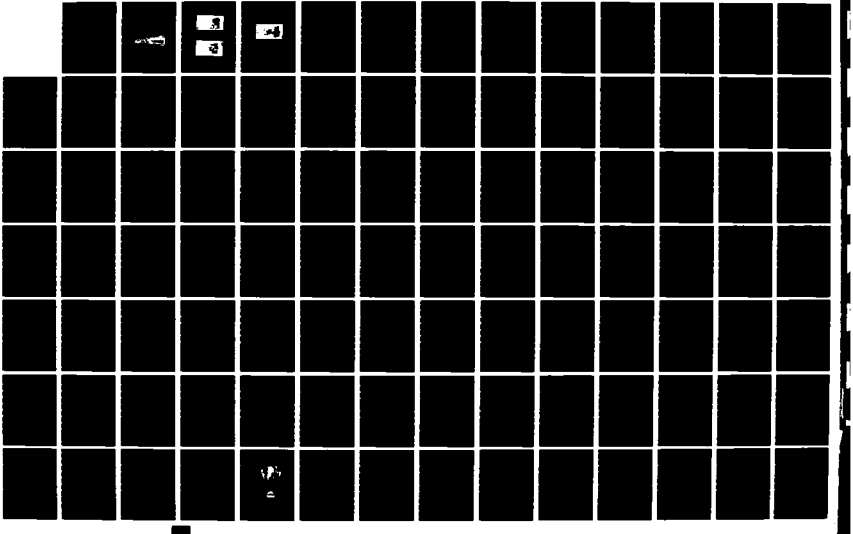
Figure 4. (continued)

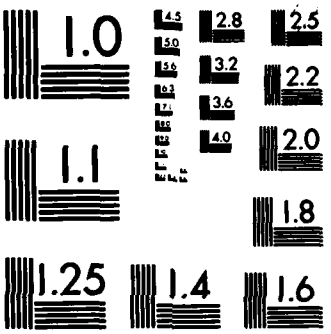


(d)

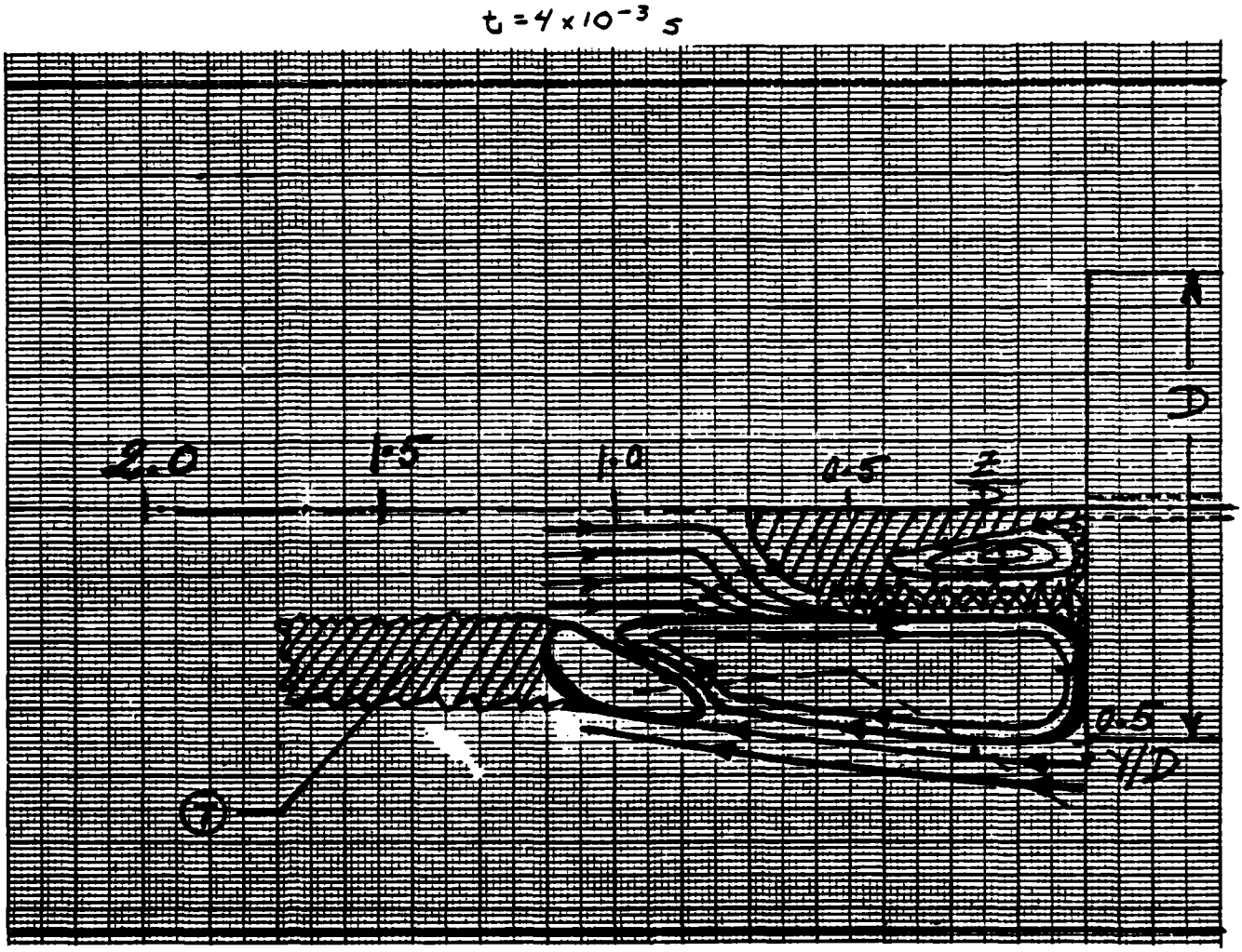
Figure 4. (continued)

AD-A130 770 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM (1982) 7/11
RESEARCH REPORTS VOLUME..(U) SOUTHEASTERN CENTER FOR
ELECTRICAL ENGINEERING EDUCATION INC S..
UNCLASSIFIED W D PEELE ET AL. OCT 82 AFOSR-TR-83-0614 F/G 5/1 NL





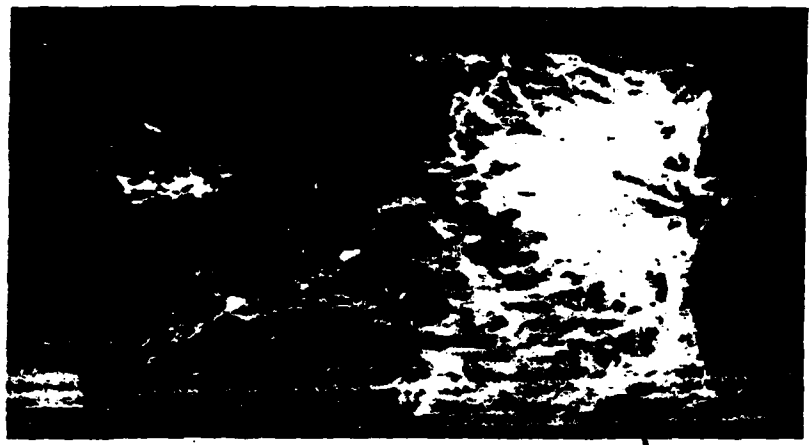
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



(e)

Figure 4. (continued)

Air flow rate = 1kg/s
Fuel flow rate = 6kg/hr



(1)



(2)

Figure 5. Photographic stills made from cine film (framing speed,4000/s)showing air indentation (vortex) growth.

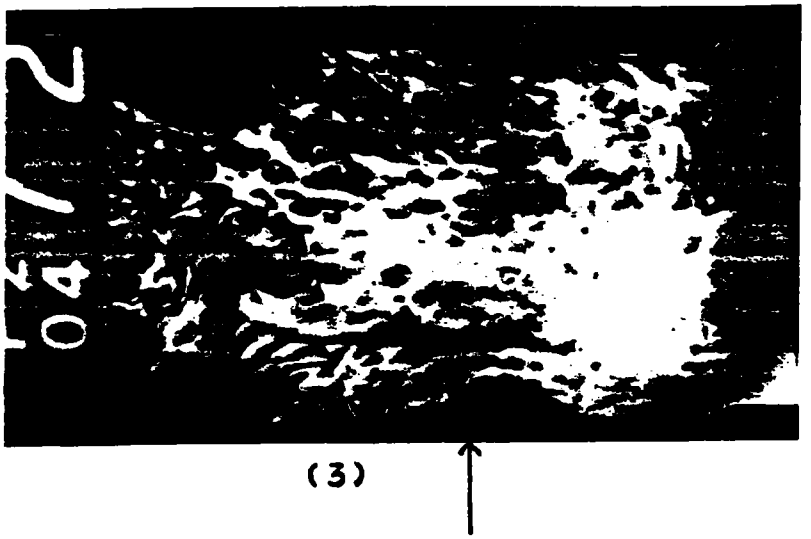


Figure 5. (continued)

= 1.0, and Y/D ~ 0.5. It is pointed out here that the outer recirculation zone boundary may fluctuate around these size limits. Air indentation or vortex, 6, is made visible by interaction of annulus air against reacting and illuminating fluid in the outer recirculation zone. Air vortex appears to grow as indicated by Fig. 4 and photographic stills, Fig. 5. Air vortex moves downstream with average estimated axial velocity of about 22.8 m/s which is close to annulus air velocity of 23.3 m/s. Assuming air vortex as an identifiable object attached to the periphery of the outer recirculation zone, estimated peripheral angular velocity of the recirculation zone is 506.5 radians/s. Air turbule growth may be attributed to its fluid heating as it moves downstream by exchange of heat energy between the relatively hot reacting fluid in the recirculation zone and vortex fluid. Some dispersion of the reacting fluid of the outer recirculating zone into the air vortex is also indicated by cine film analyses. Quasi-periodic interaction of air turbules with the outer recirculation zone fluid can be thought to augment mixing of the jet fuel with the surrounding fluid. Fig. 4 e indicates an ideal situation of an air vortex arrival at the downstream end of the outer recirculation zone. Such a situation often coincides with "slipping" of an irregularly bounded cylindrical shell of reacting fluid in the downstream direction. It is interesting to note that the ratio of air vortex to flame turbule frequency determined at location of Z/D ~ 0.7 is about two to one, and average flame turbule length being about 9 cm determined at this location. These results are specifically valid for air flow rate of 1kg/s.

Fluid evacuation of the recirculation region in terms of frequency associated flame turbules may be speculated as a mechanism to maintain time average mass conservation of the recirculating fluid spatial

region. Just after the moment of "detachment" of a flame turbule, upstream moving fluid appears to speed-up to undo the effect of fluid evacuation of the recirculation region. The upstream moving fluid is speculated to slow down as mass and thermal energy content of the recirculation region dynamically increases to a "saturation" point for an air and fuel flow rate combination. The moment the saturation point is reached, the recirculation zone is thought to be most unstable and ready to loose fluid in the form of irregularly surfaced cylindrical shells.

The view about the rotation of air indentation or vortex, 6, is as follows. Air vortices are generated from shear layer of annulus air stream in the vicinity of the bluff-body cylindrical surface. The probable rotational sense, clockwise, is in opposition to that of the outer recirculation zone, 5 . Rotational sense of the air vortex should not be necessarily taken as opposing the linear motion of fluid elements on the periphery of the outer recirculation zone. If one imagines that the downstream directional net of the downstream convection and tangential velocity of a point on the vortex periphery closest to the recirculation zone is greater or equal to the linear velocity of a fluid element on the periphery of the recirculation zone, rotating vortex does not oppose motion of the recirculation zone fluid. This explanation is in agreement with what has been observed in relevance to the dynamic behavior of the near flow field spatial region. Alternatives to the above explanation are: (a) no air vortex rotation, (b) a pair of two fluid vortices, one nearer to the air stream, rotating clockwise and the other nearer to the recirculation zone rotating counterclockwise. At the time of the writing of this report the explanation given prior to

the alternatives appears to be more appropriate. Vortex angular velocity estimated from photographic stills is 548.9 rad/s.

A comment concerning the relationship of the flow-field to the combustion process is as follows. Fuel combustion takes place in the spatial region consisting of inner and outer recirculation zones sandwiching a mixed upstream flowing fluid in between them in three dimensional space. Escape of the fuel from the near flow field region can be thought of in terms of irregularly surfaced cylindrical flame turbules slipping downstream off the flow-field region, at least, for fuel flow rates of less than 6kg/hr and air flow rate of 1kg/s.

Air Indentation (or vortex) Axial Velocity, Growth Rate, Frequency

Table 1 shows the air indentation estimated axial velocity component and growth rate data used to determine the average values. Arithmetic average values of axial velocity component and growth rate are 22.78 $\pm$ 11.4 (std. dev.) m/s, and 12.6 $\pm$ 8.96 (std. dev.) m/s, respectively. Data show the variation in axial velocity and growth rate as a function of the dimensionless axial distance, Z/D. Even at the same location respective parametric values differ for different air turbules. Variation in these parametric values could be explained in terms of spatial and temporal variations in chemical reaction rate resulting in heat release rate, and variation in fluid dynamic interaction process occurring at the interface between the reacting fluid in the recirculation zone and the relatively less hot surrounding air stream.

Table 2 shows air indentation (vortex) frequency determinations at two axial locations of Z/D = 0.36 and 0.68. Frequency at Z/D = 0.36 is 177.4 $\pm$ 8.8 (std. dev.) Hz as compared with that of 143.4 $\pm$ 9.8 (std.

TABLE 1

AIR INDENTATION OR VORTEX AXIAL VELOCITY COMPONENT
AND ITS GROWTH RATE

Air Flow Rate = 1kg/s
Fuel flow rate = 6kg/hr

Z/D, dimensionless axial distance	Axial velocity component, $V_{x,a}$ m/s	Growth rate, m/s
0.16	16.38	6.3
0.21	12.60	4.2
0.24	28.35	12.6
0.28	15.75	0
0.34	13.23	22.04
0.34	30.23	10.5
0.37	34.65	37.8
0.38	13.23	15.74
0.44	13.44	8.4
0.48	31.50	12.6
0.49	51.02	8.4
0.51	15.75	-
0.56	11.34	12.6
0.60	31.50	12.6
Average	22.78	12.6
±Std. Dev.	±11.4	±8.96

TABLE 2

AIR INDENTATION (OR VORTEX) FREQUENCIES

Frequency, Hz

Z, location of frequency determination from face of the bluff-body

	Z = 5 cm	Z = 9.5 cm
	168.00	144.00
	176.00	160.00
	176.00	122.00
	164.06	144.93
	194.29	142.86
	186.33	140.19
	173.16	138.41
	181.82	151.72
	170.37	146.67
	184.33	
Average	177.44	143.42
± Std Dev	± 8.78	±9.73

dev.) Hz at Z/D = 0.68. This implies that for every ten air indentations at the upstream axial location only about eight identations make themselves identifiable at the downstream location. This is indicative of probable air indentation coalescing phenomenon. Occurrence of the phenomenon is, perhaps, due to variation in the dynamic behavior of air indentations as they move downstream. This view is supported by the data in Table 2.

Flame Turbule Frequency and Length

Table 3 shows flame turbule frequency and length determined at an axial location of Z/D = 9.5/14 = 0.68. Flame turbule frequency of 70.6 $\pm$ 4.0 (std. dev.) Hz is less in comparison with air turbule frequency of 143.4 $\pm$ 9.8 (std. dev.) Hz at the same axial location. This translates into one slipping of "an irregularly surfaced cylindrical reacting fluid region from the recirculating spatial region" for every two air indentations appearing. Average flame turbule length is 9.0 $\pm$ .6 (std. dev.) cm. This type of information is suggestive of development of empirical mathematical expressions to predict air vortex and flame turbule frequencies, and, flame turbule length for a given air/fuel flow rate combination in a bluff-body diffusion flame combustor of the APL type. The empirically developed mathematical expressions might then be utilized to develop a mathematical "flame turbule" combustion model to predict combustion performance of a given combustor of the above said type.

Cold Flow CO_2 Jet Characteristics, and, Air Vortex and Flame Interface Location and Radial Velocity

Table 4 shows the various dimensions of cold flow CO_2 jet issuing into the combustor space for an air/fuel flow rate combination of 1kg/s/ 4kg/hr. It is appropriate to mention that the dimensions reported were

TABLE 3

FLAME TURBULE FREQUENCY AND LENGTH AT AXIAL LOCATION OF 9.5 CM FROM THE FACE OF THE BLUFF-BODY

	Frequency (Hz)	Length (cm)
	67.80	8.40
	72.95	
	70.26	8.40
	68.89	9.28
	78.05	9.80
	65.57	
Average	70.59	8.97
±Std. Dev.	+4.02	± 0.60

TABLE 4

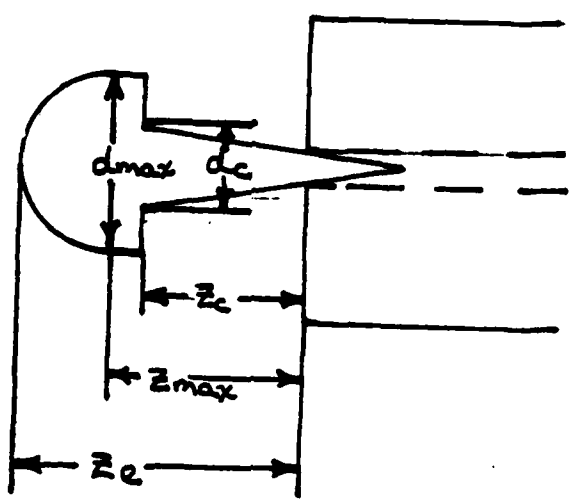
DIMENSIONS OF COLD FLOW CO_2 JET ISSUING INTO THE APL COMBUSTOR

CO_2 flow rate = 4kg/hr

Air flow rate = 1kg/s

Jet diameter at Z = 0 is 4.8mm

Set No.	Z_c, cm	d_c, cm	Z_{max}, cm	d_{max}, cm	Z_e, cm
1	6.24	1.9	6.51	4.61	9.77
2	5.70	2.44	7.60	5.43	10.31
3	5.97	2.71	7.60	5.43	10.31
4	5.70	2.71	7.60	4.89	9.23
5	5.97	2.99	7.60	5.19	10.04
6	5.70	2.17	7.60	5.43	10.59
Mean	5.88+.20	2.49	7.42	5.16±.31	9.99±.45
±Std. Dev.		±.37	±0.41		



deduced from a shadowgraphic film recorded by means of Mn-laser with average firing frequency of 6563/s. CO_2 jet diameter spreads from 0.48 cm at Z/D = 0 to 2.49 cm at Z/D = 5.88/14 = 0.42 in a linear manner. Downstream of Z/D = 0.42, jet spreads in to "mushroom". Jet mushrooming phenomenon is explained in terms of the backward flow and the jet fluid interaction, and the influence of the dynamic behavior of the jet surrounding flow-field. Jet mushroom maximum diameter is 5.16 cm at Z = Z_{max} = 7.42 cm; therefore, (jet mushroom maximum dia.)/(jet cone maximum dia.) is 2.0. Maximum jet length is about 10.0 cm. Such type of information regarding fuel jet is useful in developing predictive combustion model in light of involvement of the fuel jet surface and volume in reactant species mixing and combustion processes.

Table 5 shows the air vortex and flame interface location and radial velocity towards the combustor centerline. Variation in Yint/D as a function of Z/D may be envisioned as "flame surface wrinkling" of the reactive zone. Yint/D fluctuates within 0.31 - 0.48 for Z/D variation from 0.16 to 0.60. Variation in the interface radial velocity towards the centerline as function of Z/D is also observed.

TABLE 5

RADIAL VELOCITY COMPONENT, $V^{+}_{y,t}$, INTERFACE BETWEEN THE OUTER RECIRCULATION ZONE REACTING FLUID AND AN AIR INDENTATION OR VORTEX

Air flow rate = 1kg/s
Fuel flow rate = 6kg/hr

Z/D	Yint/D	$V_{y,t}$, m/s
0.16	.48	0.00
0.21	.47	-1.68
0.24	.47	-6.93
0.28	.45	-7.56
0.34	.39	-0.63
0.34	.44	-5.88
0.37	.43	-4.41
0.38	.38	-5.04
0.44	.35	-6.72
0.48	.40	-10.71
0.49	.42	-2.10
0.51	.33	-3.78
0.56	.31	-8.19
0.60	.38	-1.26
Average		-4.64
$\pm$ Std. Dev.		$\pm$ 3.10

+ Movement of the interface along radius towards the centerline is considered negative

TABLE 6

Recirculation Zone Rotational

Velocity determination

Air flow rate = 1kg/s
Fuel flow rate = 6kg/hr

Vortex #	Rotational velocity (rad/s)
1	546.87
2	477.06
3	314.16
4	570.14
5	624.0
Average	506.45
Std. Dev.	$\pm$107.1

V. Summary of Main Conclusions

1: Dynamic behavior of the near-region combusting flow field in the wake of the bluff-body appears to be resulting from temporal and spatial interactions among the jet fuel, upstream moving fluid, and annular air stream. Such interactions lead to an interesting flow-field as shown in Figure 4 for air flow rate of 1kg/s and fuel flow rate of 6kg/hr or less.

2. "Outer recirculation zone" dimensionless size limits appear to extend to about 1.0 and 0.5 in axial and radial directions, respectively, for air flow rate of 1kg/s and fuel flow rate of 6kg/hr or less. Average peripheral rotational velocity of the recirculation zone is 506.5 rad/s.

3. Fuel escaping from combustion in the near flow field spatial region, in the wake of the bluff-body, can be thought of in terms of irregularly surfaced flame cylindrical shells that slip off the flow field region quasi-periodically.

4. Ratio of air vortex/flame turbule or shell frequency is about 2/1 for the flow rate combination investigated (Air flow = 1kg/s, fuel flow = 6kg/hr).

5. Air indentation or vortex interacting with the reacting, heat generating, and light emitting fluid has estimated average axial velocity of 22.8 m/s and rotational velocity of 548.9 rad/s.

6. The internal cold flow conical jet appears to spread linearly with the axial distance and then changes into instable "mushroom". The jet fluid appears to be pulled into the surrounding environment as partially indicated in Figure 4.

VI. Recommendations

Although by means of the information generated from cine and high speed shadowgraphic photography it has been possible to develop a partial picture of the dynamic behavior of the near region flow field in the wake of the bluff-body in the APL combustor, more quantitative information regarding time variant system variables such as velocity and pressure throughout the entire flow field would be needed to complete the picture. Truly, one needs to know instantaneous values of the above mentioned variables simultaneously at a number of points in the spatial region in the wake of the bluff-body. Local instantaneous information in the entire flow field should be gathered as function of time for an isothermal cold flow under investigation. Because for a combusting flow dynamic behavior would be different, at least in magnitude of the values of behavior related parameters if not in gross trends, simultaneous spatial time variant quantitative information regarding velocity, pressure, and, for maximum usefulness of the generated data, additional information regarding temperature and species concentrations should be experimentally acquired. A comprehensive data acquisition scheme is suggested as follows. For each combination of air and fuel flow rates spatial region extending from Z/D = 0. to 1.5 and, Y/D = 0. to 0.7 is proposed for probing with 1/2 cm distance between adjacent points. For each of two air flow rate conditions of 1 and 2kg/s, fuel flow rate conditions to be investigated are 2,4,6 and 8,10,12 kg/hr. In addition, cine films for combusting flows and high speed shadowgraphic films for isothermal CO_2 and air jet combinations would be useful. After gathering such experimental information one would be in a better situation to gain deeper insight and develop a more sophisticated explanation of the dynamic behavior of the flow field, and, hence, of

the combustion process.

REFERENCES

1. Roquemore, W.M., Bradley, R.P., Stutrud, J.S., Reeves, C.M., and Krishnamurthy, L., "Preliminary Evaluation of a Combustor for Use in Modeling and Diagnostic Development," ASME Publication 80-GT-93, March, 1980.

2. Sandhu, S.S., "Study of Dynamic Behavior of a Bluff-body Diffusion Flame in the APL Combustion Tunnel Facility," 1981 USAF SCEE Summer Faculty Research Program, Final Report, Contract No. F49620-79-0038.

3. Roquemore, W.M., Britton, R.L., and Sandhu, S.S., "Investigation of the Dynamic Behavior of a Bluff-body Diffusion Flame Using Flame Emission," presented at the AIAA 20th Aerospace Sciences Meeting, January, 1982.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

PERFORMANCE DEMAND AND CNS DEPRESSANT STRESSOR EFFECTS ON SPATIAL ORIENTATION INFORMATION PROCESSING

Prepared By:	Dr. Robert E. Schlegel
Academic Rank:	Assistant Professor
Department and University:	School of Industrial Engineering University of Oklahoma
Research Location:	USAF School of Aerospace Medicine, Crew Technology Division, Crew Performance Branch
USAF Research Colleagues:	Dr. John F. Courtright Dr. William F. Storm
Date:	September 7, 1982
Contract No:	F49620-82-C-0035

PERFORMANCE DEMAND AND CNS DEPRESSANT STRESSOR EFFECTS ON SPATIAL ORIENTATION INFORMATION PROCESSING

by

Robert E. Schlegel

ABSTRACT

An evaluation of spatial orientation information processing was performed using the Manikin Task. The specific objectives of the study were to measure the speed vs. accuracy tradeoff characteristics of the task and to assess performance on the task under the influence of ethyl alcohol.

Five male subjects were extensively trained in the performance of the task. Following training, the subjects performed the task at various presentation rates (workload demand) and under baseline vs. alcohol conditions (CNS depressant). Preliminary data analysis revealed that the task is in fact sensitive to both workload demand and the presence of alcohol. The speed-accuracy tradeoff function shifts in the direction of both decreasing speed and decreasing accuracy. Recommended extensions of the research in terms of further data analysis and examination of additional research questions are provided.

ACKNOWLEDGEMENTS

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the rare opportunity to spend a very productive and enjoyable summer with the Crew Performance Branch at the USAF School of Aerospace Medicine, Brooks AFB, Texas. In addition, he would like to thank the entire staff of the branch and the many other individuals who provided substantial support during the relatively short research period.

Specifically, appreciation is expressed to Dr. John F. Courtright who provided the initial suggestion for the research topic and continued to provide technical, personnel and moral support in the midst of a hectic reorganizational effort. The author is also indebted to Dr. William F. Storm for his suggestions on improving the experimental methodology and for his constant encouragement and to Dr. Loren G. Myhre for his helpful discussions and invaluable assistance.

A final note of thanks is offered to those individuals who agreed to participate as experimental subjects during the many data collection trials.

I. INTRODUCTION:

An important goal of human factors research is to develop safe, efficient, cost effective systems by utilizing available knowledge of task requirements, environmental and workload constraints and human limitations. The task of piloting high-performance jet aircraft requires a multitude of complex psychomotor skills, high rates of information processing and high-speed decision making. Additional stressors place even greater demands on the capabilities of Air Force flight crews.

In analyzing the task of piloting aircraft, one obviously vital information processing component involves the ability to perform spatial orientation activities. It is essential that the pilot and other crew members maintain the ability to differentiate right vs. left and up vs. down based on available visual cues often under severe conditions of disorientation and exposure to various types of stressors. Furthermore, it is important to investigate the process of spatial orientation information processing in the research laboratory to find appropriate means of measuring this capability and to evaluate its sensitivity to different stressing agents. Of particular interest are the effects resulting from varying workload demands and certain chemical defense drugs which act as central nervous system (CNS) depressants.

The Manikin Task was originally conceived by Benson and Gedye (1) at the Royal Air Force (RAF) Institute of Aviation Medicine (IAM), Farnborough, U.K. and employed as a complex reaction time task related to the ability of pilots to orient themselves with respect to an external visual reference. The task consists of a sequence of sketches of human figures (manikins) in various orientations. The orientation categories consist of (1) front-erect, (2) front-inverted, (3) back-erect and (4) back-inverted.

For each presentation, the subject is required to identify in which hand the manikin holds a particular shape (circle or square) identical to one presented below the sketch. This requires the subject to execute some form of mental rotation of the figure about one or more axes to obtain the proper reorientation and discriminate right from left.

The original implementation of the task employed 35mm slides to present the series of orientations. Improved versions have been developed using computer generated CRT displays. A sophisticated version which provides for manipulation of several task parameters has been recently implemented on a PDP 11/34 computer system.

A comprehensive assessment of the utility of the Manikin Task has been of interest to researchers at the USAF School of Aerospace Medicine's Crew Performance Branch. Through a series of experiments, it has been shown that the

task satisfies many of the necessary criteria to be considered a useful laboratory measure. In summary:

1. subjects can learn the task and acquire plateau performance within 4 to 10 training sessions each of 30 minutes duration

2. the rate of acquisition of plateau performance is essentially independent of the training schedule, occupational level and age of the subject

3. the task has a high level of differential stability meaning that individual performance remains essentially constant over time (Reader et al.: 2).

In addition, the task has been shown to be sensitive to mild hypoxia (Benel and Storm: 3) but insensitive to changes in head temperature under hyperthermic conditions (Nunneley, Reader, and Maldonado: 4).

A logical continuation in the assessment of the Manikin Task is to evaluate its validity with respect to other human stressors with the added intent of gaining a better understanding of the spatial orientation process.

II. OBJECTIVES:

The general objective of the summer research effort was to evaluate cognitive processing activity related to spatial orientation tasks. The task selected for this evaluation

was the Manikin Task. Two general categories of human stressors were investigated in terms of their effects on performance of the task, Workload demand stressors and CNS depressant stressors. This provided two specific objectives for the project:

1. to measure the speed vs. accuracy tradeoff characteristics of the task (Workload Demand Stressor)

2. to assess performance on the task under the influence of ethyl alcohol (CNS Depressant Stressor)

It was also believed that the knowledge gained from collecting additional data on the task would lead to recommendations for improved task administration methodologies. These would include various procedural aspects, stimulus presentation rates and certain adaptive test variations.

III. SPEED-ACCURACY TRADEOFFS RELATED TO ALCOHOL

With respect to many psychomotor tasks, it has been proposed that the inconsistency of previous results concerning the effects of alcohol on reaction time (RT) may be related to possible tradeoffs between speed and accuracy (Jennings et al.: 5). In the majority of information processing studies in which RT is utilized as the primary criterion, subjects are usually encouraged to respond "as rapidly and accurately as possible." The RT and error rate values thus represent a compromise between the incompatible

demands for maximum accuracy and minimum RT. This tradeoff may be represented by the function illustrated in Figure 1.

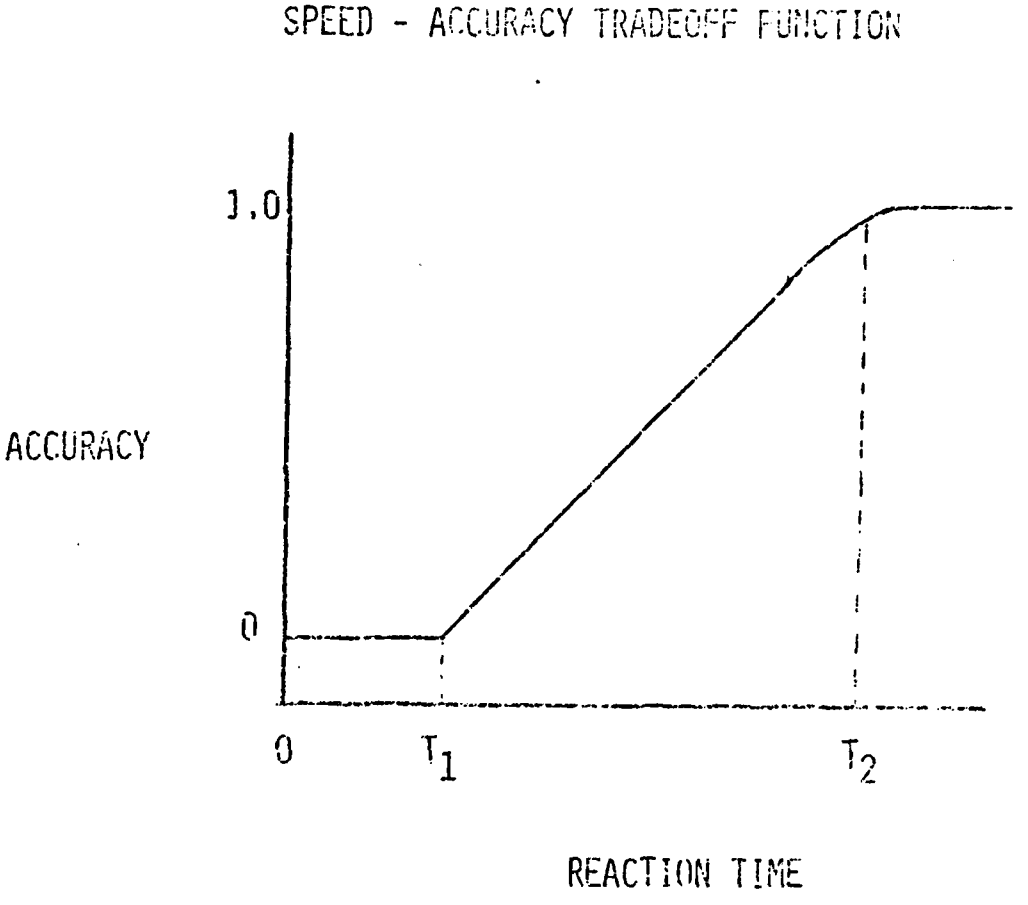


Figure 1. Idealized speed-accuracy tradeoff function (Wood and Jennings: 6)

Unfortunately, the particular compromise between speed and accuracy that is adopted for any given experimental trial is a function of several variables including the actual experimental conditions under investigation. It is therefore difficult to derive any valid conclusions from experiments in which the change in subject criteria is not identified and measured.

Although perfect accuracy is rarely achieved in RT studies, relatively high levels can often be attained. Reported data confirms that accuracies greater that 90% are common (Wood and Jennings: 6). However, these accuracy levels may be unrealistic in view of the fact that task demands often do not allow sufficient time for a highly accurate response. Also, an examination of the tradeoff function in Figure 1 reveals that very small changes in error rate at high levels of accuracy are associated with large changes in RT. Again, this makes it difficult to determine if significant differences in measured reaction times are due to experimental conditions or merely related to slight shifts in the criterion point.

One proposed solution to these problems is to utilize the complete tradeoff function as a measure of information processing performance in which changes in bias for speed and accuracy can be directly assessed. This approach allows one to distinguish a shift in the subject's speed-accuracy criteria from a possibly more significant change in processing efficiency independent of a shift in criteria. Processing efficiency differences can be easily identified by changes in either the slope or RT intercept of the tradeoff function.

This procedure has been employed in at least two studies evaluating the effects of alcohol on speed-accuracy tradeoffs. In an experiment by Jennings et al. (5), five

levels of alcohol were examined ranging from 0.00 to 1.33 ml/kg of body weight. Increasing doses of alcohol produced a progressive decrease in the slope parameter of linear equations fit to the speed-accuracy data, but did not significantly alter the intercept of the functions with the RT axis. Thus, alcohol reduced performance efficiency by decreasing the rate of growth of accuracy per unit time. Similar results were obtained in a study by Rundell and Williams (7) who examined alcohol levels from 0.0 to 1.0 g/kg of body weight.

In both studies, deadline procedures were used to force subjects to respond within various time limits. The time limits were selected to provide accuracies ranging from chance levels to near perfection. These procedures alter the workload demand such that the shorter time limits create extreme time-stressed conditions.

A similar approach was employed in the current effort to generate the speed-accuracy tradeoff functions for the Manikin Task performed under baseline and alcohol conditions.

IV. EXPERIMENTAL METHODOLOGY

Subjects

Five male subjects, ages 18 to 49, participated in the investigation which consisted of initial training sessions, testing sessions to derive the speed-accuracy tradeoff

function under baseline conditions and testing sessions employing alcohol. All subjects were right-handed and had normal (20/20 corrected) vision. All were light to moderate social drinkers averaging 4 to 8 drinks (mixed drinks, beers or glasses of wine) per week. Each subject performed the task under both the baseline and alcohol conditions. The voluntary informed consent of the subjects was obtained in accordance with AFR 169-3.

Task

Each subject was instructed to observe a 25cm CRT display (Digital Equipment Corporation Model VR17LC) located approximately 1 meter in front of him. For each stimulus presentation, the subject was to mentally reorient the manikin and decide in which hand the manikin held a shape identical to the shape displayed at the bottom of the screen. A response was registered by pressing either the right or left button on a subject panel, always using the same finger of the right hand.

In the initial training sessions, each manikin remained on the screen for 2 seconds with a 1-second pause between presentations. A series of 96 manikins was presented in a sequence followed by 2 minutes of rest. Four sequences were grouped to form a single testing session lasting approximately 25 minutes. All subjects participated in a minimum of 8 training sessions. Previous results have shown that an average of 6 sessions are needed to achieve stable perfor-

mance on the task. The criteria for terminating training required a deviation of no more than ± 5% of the mean reaction time (RT) of the previous 2 sessions.

Following training, the amount of time that the manikin remained on the screen was reduced to levels ranging from 400 to 1000 msec. The pause between presentations was increased accordingly to maintain a 2-second interval between the start of successive presentations. All other variables remained the same as during the training sessions. Subjects were instructed to respond "at of before the instant the manikin leaves the screen." This provided the deadline conditions essential for measuring subject performance over a wide range of speed and error rates.

Alcohol

In addition to the baseline condition, a single alcohol dose consisting of 0.5 to 0.75 grams of pure alcohol per kilogram of body weight was administered to each subject. The actual dose depended on the body composition of the individual subject and was the calculated dose required to raise the subject's blood alcohol concentration (BAC) to a level of 0.08%. Subjects were instructed not to consume alcoholic beverages on the evening before the alcohol testing trial. Testing sessions were conducted in the morning prior to the consumption of any food or drink other than water.

The equivalent volume of 86 proof bourbon was combined with water in the ratio of two parts water to one part alcohol. Subjects were allowed 15 minutes to consume the beverage. Following consumption, blood alcohol levels were indirectly measured using an Intoxilyzer, Model 4011 manufactured by CMI Incorporated according to the procedures of Dubowski (8) and Spector (9).

Testing was initiated approximately 30 minutes after consumption of the alcohol. To sustain peak BAC levels and thereby circumvent problems associated with differential performance on the ascending and descending limbs of the blood alcohol function, a maintenance dose of 0.05 g/kg of body weight was administered every 20 minutes (Rundell and Williams: 7). Following the experimental task, the subject was monitored until his BAC was at or below 0.005% and then released.

Procedure

For the alcohol evaluation phase of the study, each subject entered the laboratory in the early morning and performed three baseline testing sessions, one each under the 1000, 700 and 400 msec deadline conditions. The subject was then weighed and the appropriate dose was calculated. The alcohol was then administered and sufficient time elapsed for the subject to attain the desired BAC level. The subject then performed three additional testing sessions, again at 1000, 700 and 400 msec. Only one subject was tested per

day during the alcohol phase. Five weeks of data collection were required for the entire investigation. All data was collected using a PDP 11/34 computer system and subsequently transferred to tape for further analysis.

V. RESULTS

The results of the research effort can be divided into three categories: (1) training phase, (2) speed-accuracy tradeoff functions and (3) alcohol evaluation phase. Due to the massive amounts of data that were collected, only a superficial analysis has been completed at this time.

With respect to the training phase, it was evident that all subjects attained stabilized performance within the first eight sessions with some subjects requiring only four to five sessions. It was also evident that without the deadline conditions, high accuracy rates (95 to 99%) were attained even during the initial stages of training. The data also confirmed a previously discovered dependency of reaction times on specific manikin orientations.

The deadline procedures used in the second phase of the study successfully produced the desired shift in accuracy and reaction time. For one subject, the average values varied from 99% accuracy at 655 msec for the 1000 msec deadline to 97% accuracy at 601 msec for the 700 msec deadline to 71% accuracy at 519 msec for the 400 msec deadline. Similar results were observed for the other subjects.

Further analysis will isolate the relationship between speed-accuracy and specific orientations.

With respect to the final phase of the investigation, the alcohol administration procedures generated the desired BAC levels. In addition, a speed-accuracy tradeoff shift did occur. This can be observed by examining the data for the above mentioned subject under the alcohol condition. The average values varied from 96% accuracy at 723 msec for the 1000 msec deadline to 93% accuracy at 630 msec for the 700 msec deadline to 57% accuracy at 533 msec for the 400 msec deadline. Again, further analysis is required to examine other relationships in the data.

The major general results of the research study were that:

1. speed-accuracy tradeoff functions can be generated for the Manikin Task, thus verifying that the task is sensitive to workload demands

2. these functions <u>are</u> influenced by the presence of alcohol.

In addition, several observations were made concerning the manner in which subjects accomplished the Manikin Task. Evaluation of these observations will be accomplished by a more detailed analysis of the data.

VI. RECOMMENDATIONS

Several recommendations exist for future analysis of the data and extension of the research activity. First, a thorough analysis of the speed-accuracy tradeoff functions needs to be performed. Wood and Jennings (6) have discussed several transformations that may be applied to the accuracy data. These include various measures of information transmission and signal detection theory. When applied to both the baseline and alcohol data, it should be possible to obtain a better understanding of the different states.

Second, there were various differences in both RT and error rate which appeared to be dependent on the particular manikin orientation, square-circle pattern, right vs. left preference and other factors. For each of the almost 10,000 trial presentations made during the data collection period, complete information was recorded on these various factors. Detailed analysis of this data should lead to some interesting discoveries concerning spatial orientation tasks.

Finally, examination of the individual subject alcohol response curves will contribute to the author's understanding of alcohol ingestion and metabolism rates.

In terms of extending the research activity, an important question arose as to whether trained subjects continued to employ a particular mental orientation algorithm or relied merely on pattern recognition. The answer to this

vital question could greatly aid the understanding of the spatial orientation process. Additionally, the use of EEG analysis and the measurement of evoke potentials as they correlate with specific orientations could help establish the underlying process.

Since different orientations elicit different RT and error rate values, a study should be made of the feasibility of varying the workload demand by changing the relative percentage of each orientation. This could also lead to the development of better adaptive test methodologies in which the difficulty of the task is based on the subject's current level of performance.

Finally, since alcohol has been shown to have a significant effect, a full range of alcohol levels should be investigated followed by an examination of other stressors including extended durations, circadian rhythm changes and chemical defense agents.

REFERENCES

1. Benson, A.J. and Gedye, J.L., Logical Processes in the Resolution of Orientational Conflict, RAF Institute of Aviation Medicine, Rep. No. 259, Ministry of Defense (Air), 1963.

2. Reader, D.C., Benel, R.A. and Rahe, A.J., Evaluation of a Manikin Psychomotor Task, SAM-TR-81-10, USAF School of Aerospace Medicine, Brooks AFB, Texas, May, 1981.

3. Benel, R.A. and Storm, W.F., "Performance Decrement at Cabin Altitudes: A Replication," Proceedings of the 25th Annual Meeting of the Human Factors Society, Rochester, New York, October, 1981.

4. Nunneley, S.A., Reader, D.C. and Maldonado, R.J., "Head-Temperature Effects on Physiology, Comfort, and Performance During Hyperthermia," Aviation, Space, and Environmental Medicine, 53(7), pp. 625-628, 1982.

5. Jennings, J.R., Wood, C.C. and Lawrence, B.E., "Effects of Graded Doses of Alcohol on Speed-Accuracy Tradeoff in Choice Reaction Time," Perception and Psychophysics, 19(1), pp. 85-91, 1976.

6. Wood, C.C. and Jennings, J.R., "Speed-Accuracy Tradeoff Functions in Choice Reaction Time: Experimental Designs and Computational Procedures," Perception and Psychophysics, 19(1), pp. 92-101, 1976.

7. Rundell, O.H. and Williams, H.L., "Alcohol and Speed-Accuracy Tradeoff," Human Factors, 21(4), pp. 433-443, 1979.

8. Dubowski, D.M., "Measurement of Ethyl Alcohol in Breath," In: Sunderman, L.F. (Ed), Laboratory Diagnosis of Diseases Caused by Toxic Agents, Warren H. Green Inc., St. Louis, Missouri, 1970.

9. Spector, N.H., "Alcohol Breath Tests: Gross Errors in Current Methods of Measuring Alveolar Gas Concentrations," Science, 172, pp. 57-59, 1971.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

METAL ORGANIC CHEMICAL VAPOR DEPOSITION-EVALUATION OF CURRENT AND POSSIBLE FUTURE CHEMICAL SYSTEMS AS RELATED TO THE FORMATION OF GALLIUM ARSENIDE

Prepared by:	Dr. Eugene P. Schram
Academic Rank:	Associate Professor
Department and University:	Department of Chemistry The Ohio State University
Research Location:	Avionics Laboratory, Electronics Research Branch, Wright Patterson Air Force Base
USAF Research Colleague:	Dr. Philip E. Stover
Date:	August 15, 1982
Contract No:	F49620-82-C-0035

METAL ORGANIC CHEMICAL VAPOR DEPOSITION-EVALUATION OF CURRENT AND POSSIBLE FUTURE CHEMICAL SYSTEMS AS RELATED TO THE FORMATION OF GALLIUM ARSENIDE

by

Eugene P. Schram

ABSTRACT

Current chemical systems utilized for the formation of gallium arsenide, via MOCVD, have been compared. The highest quality product has apparently resulted from the employment of gallium triethyl and arsine. However, care must be exercised in comparing different chemical systems because the purity of all reactants is critical to the production of high mobility GaAs.

Potential new molecular sources for gallium are suggested, as related to MOCVD, and include: $GaH_3 \cdot N(CH_3)_3$, $GaH_3 \cdot NH_3$, and several polyfluoroalkyl derivatives of gallium.

Acknowledgement

The author wishes to express his sincere appreciation to the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for the opportunity to become acquainted with the extremely interesting research activities carried out within the Electronics branch of the Avionics laboratory at Wright Patterson Air Force Base. The enthusiasm expressed and friendly exchange of information was much appreciated, and I wish to thank particularly Drs. Phil Stover, Cole Litton, Elgene Nichols, Don Reynolds and Peter Colter.

I. INTRODUCTION:

The formation of gallium arsenide, GaAs, via the thermally induced vapor phase decomposition of gallium alkyls, is usually considered to begin with Manasevits' report on the reaction between gallium trimethyl, $Ga(CH_3)_3$, and arsine, AsH_3.[1,2] This general type of process is commonly referred to as MOCVD, metal organic chemical vapor deposition: This report is not intended to review exhaustively the literature on MOCVD, as related to GaAs, but to draw attention to those chemical systems which have resulted in relatively high purity GaAs with good surface morphology. However, one should note, regarding cited references, that usually only the best results are discussed, i.e. high mobiltiy GaAs. In addition to the obvious reaction parameters such as reactor design, hot wall or cold wall reactors, flow rates, ratios of As/Ga in the reaction gases, heating rates, reaction temperature and methods of heating, one finds in recent reports advantages of carrying out the decomposition at sub-atmospheric pressures. Finally, various workers have employed different chemical species as the sources for Ga and As.

II. OBJECTIVES:

MOCVD, as related to GaAs, has the demonstrated potential to afford, on a production bases, thin films of high purity at initial costs well below that of molecular beam epitaxy, MBE. Furthermore, the insitu passivation of GaAs thus formed would eliminate the usual cleaning and polishing steps associated with GaAs formed by the chloride process. This report is intended to draw attention to those reports which indicate the greatest potential for the formation of high quality GaAs, and to review and recommend new chemical systems that may afford high purity GaAs.

III. FORMATION OF GAAS VIA MOCVD:

Ideally the chemical sources of gallium and arsenic should be species which are volatile at 25° (thus eliminating the need for heating large portions of the apparatus in order to vaporize the reaction components), and allow, as a consequence of thermal decomposition, only the deposition

and incorporation of Ga and As. Furthermore, it is very important that both the gallium and arsenic compounds are high purity, i.e. > 99.999%. The most common source for arsenic is the volatile and poisonous gas, arsine, AsH_3. Regarding the latter, the thermal decomposition by-product, H_2, is not incorporated into GaAs and is easily removed from the reaction system.

Sources of gallium have included galliumtrimethyl, $Ga(CH_3)_3$,[1,6] galliumtriethyl, $Ga(C_2H_5)_3$,[1,7,8] and diethylgallium chloride, $(C_2H_5)_2GaCl$.[9,10] In a related manner, the 1:1 Lewis acid-base adducts, $(C_2H_5)_2ClGa{\cdot}As(C_2H_5)_3$[11] and $(C_2H_5)_2ClGa{\cdot}As(CH_3)_3$[12] have been subjected to pyrolysis. In the last system the arsenic source is the incorporated Lewis base.

With regard to the MOCVD system, $Ga(CH_3)_3$-AsH_3, Bass[4] reported GaAs formation in the temperature range 650-760°C with carrier concentration ~5 x $10^{15}cm^{-3}$ and a room temperature Hall mobility of 6260 $cm^2v^{-1}s^{-1}$ (n-type). In a similar fashion Seki[6] found n~10^{16} cm^{-3} with a room temperature Hall mobility, μ, ~6000 $cm^2v^{-1}s^{-1}$ for GaAs produced between 500-850°C, and 1 atm. More recently Manasevit[3] produced GaAs, at 1 atm. and in the temperature range 600-625°C, with μ_{77} ~102,000 $cm^2v^{-1}s^{-1}$, and N_d-N_a ~3 x $10^{14}cm^{-3}$. In a similar manner, he found when this synthesis was carried out at 70 torr the GaAs, produced in the temperature range 575-600°C, exhibited μ_{77} ~125,000 $cm^2v^{-1}s^{-1}$ with N_d-N_a ~5 x $10^{13}cm^{-3}$. As indicated in the latter communication, superior quality GaAs appears to be favored by sub-atmospheric reaction conditions. However, care must be exercised in comparing the quality of GaAs produced under different reaction conditions because the purity of starting materials alone, under identical reaction conditions, has been shown to greatly influence the quality of the final reaction product, GaAs. That is, every aspect of the MOCVD process must be considered in judging which reaction parameters are associated with the highest quality product in a specific apparatus. In addition to those parameter previously indicated, substrate preparation is very important as is its subsequent pre-pyrolysis cleaning-oxide removal-in the reaction system. Until MOCVD is better understood every possible reaction parameter must be considered an important variable and control exercised to best afford high quality GaAs.

An alternate volatile source for Ga is $Ga(C_2H_5)_3$, first utilized by Manasevit[2], and later expanded on by Seki[8], while more recently this reagent has been developed by Change, et. al.[7] Seki, et. al., working at atmospheric pressure, obtained GaAs with an electron concentration of $7 \times 10^{13} cm^{-3}$ and μ_{77} ~120,000 $cm^2v^{-1}s^{-1}$. Furthermore, carrier concentrations could be changed in the range from 10^{12} to ~$10^{16} cm^{-3}$ by changing the AsH_3 to $Ga(C_2H_5)_3$ ratio. Galliumtrimethyl has been utilized at relatively low pressure, 20 to 25 torr, and in the temperature range 660-750°C, affording GaAs with an electron concentraion of $<10^{13} cm^{-3}$ and μ_{25} =7000 $cm^2v^{-1}s^{-1}$ (T=660°C, P=20 torr). One possible advantage of $Ga(C_2H_5)_3$ over $Ga(CH_3)_3$ is the apparent lower carbon impurity concentration in the resulting GaAs. This result may be rationalized by a possible mode of decomposition for $Ga(C_2H_5)_3$, i.e. β elemination with the overall reaction represented by eq.1, with subsequent

$$Ga(C_2H_5)_3 \longrightarrow "GaH_3" + 3CH_2{=}CH_2 \qquad (1)$$

decomposition of the intermediate, gallane, "GaH_3", eq. 2.

$$"GaH_3" \longrightarrow Ga + 3/2H_2 \qquad (2)$$

The proposed β elimination is well known for aluminium derivatives, eg. $Al(C_2H_5)_3$[13] undergoes thermal decomposition as represented by eq. 3.

$$Al(C_2H_5)_3 \longrightarrow HAl(C_2H_5)_2 + CH_2{=}CH_2 \qquad (3)$$

The formation of an olefin eq. 1, serves as a relative low energy pathway for carbon elimination from $Ga(C_2H_5)_3$ as contrasted by CH_4 elimination from $Ga(CH_3)_3$. In the latter case α elimination is also possible, under the reaction conditions, with production of the reactive carbene, CH_2. Reaction of the latter with the GaAs surface followed by thermal decomposition, to afford H_2, provides a reasonable mechanism for carbon incorporation in GaAs. Indeed, Zaouk has recently reported the formation of $H_2C{=}CH_2$ during the pyrolysis of $ClGa(C_2H_5)_2 \cdot As(CH_3)_3$.[12]

Lewis acid-base adducts between Ga and As, respectively, have been employed as volatile precursors to GaAs formation, eg, $(C_2H_5)_2ClGa\cdot As(C_2H_5)_3$[11] and $(C_2H_5)_2ClGa\cdot As(CH_3)_3$[12]. The relative Lewis base strength, towards GaR_3, follow the order N>P>As.[14] The employment of a chloro-derivative, of an alkyl gallane, increases the Lewis acidity of gallium via the negative inductive affect of Cl thereby increasing the stability of the resulting adduct in the vapor phase prior to thermolysis.

The application of adducts between Ga and As species, for the formation of GaAs, has not been particularly successful. For example, Zaouk, et.al.,[12] observed only polycrystalline GaAs resulting from the pyrolysis of $Cl(C_2H_5)_2Ga\cdot As(CH_3)_3$, between 525 and 600°C; furthermore no deposition of GaAs took place above 600°C. In a similar fashion whiskers of GaAs were reported to result from the utilization of $Cl(C_2H_5)_2Ga\cdot As(C_2H_5)_3$[11] and as has been discussed,[12] the ratio of As/Ga, in the vapor phase reactants, must be greater than one to afford epitaxial GaAs. Therefore the failure of adduct chemical systems, toward satisfactory GaAs growth, is not surprising.

IV. POTENTIAL ADDITIONAL SOURCES OF GALLIUM AS APPLIED TO MOCVD:

(A) Gallium Hydride derivatives.

As previously mentioned, the ideal gallium species is one that may be obtained in high purity, is volatile at 25°C, and may be thermally decomposed with complete volatilization of the ligand moieties, ie, (those other atoms or functional groups chemically associated with gallium). The simplest compound that meets these criteria is gallane, "GaH_3". Gallane is not stable as a monomeric species, however, the stabilized adduct may be prepared, as represented by eq.4 [15,16] while the precursor species, $LiGaH_4$, results from the synthesis represented by eq. 5.

$$LiGaH_4 + (CH_3)_3N\cdot HCl \longrightarrow GaH_3\cdot N(CH_3)_3 + LiCl + 1/2H_2 \qquad (4)$$

$$4LiH + GaCl_3 \longrightarrow LiGaH_4 + 3LiCl \qquad (5)$$

The adduct, $GaH_3\cdot N(CH_3)_3$, exhibits a vapor tension of 3 torr at 25° and melts

at 70.5°C. Other adducts of "GaH_3" are known, however, they are less stable, eg., $GaH_3 \cdot O(C_2H_5)_2$[18], and $GaH_3 \cdot S(CH_3)_2$.[19]

With regard to gallane, "GaH_3", originally this species was reported to be stable as the dimer, digallane, $[GaH_3]_2$, exhibiting at mp at -21.4°C, vapor tensions of 2.5 torr at 0°C and 49.1 torr at 54°C, bp. calcd. 139°C with an indicated decomposition temperature of 130°C.[20] The reported formation of Ga_2H_6 involved treatment of tetramethyldigallane, $Ga_2(CH_3)_4H_2$ with $N(CH_3)_3$ with the overall reaction represented by eq. 6.[20]

$$3Ga_2(CH_3)_4H_2 + 3N(CH_3)_2 \rightarrow Ga_2H_6 + 3Ga(CH_3)_3 \cdot N(CH_3)_3 + Ga(CH_3)_3 \qquad (6)$$

Subsequently Parry[19], et al tried to duplicate Wiberg's preparation of Ga_2H_6, but were unsuccessful. Parry's study was very thorough with the conclusion that Ga_2H_6 is not stable as a molecular species at room temperature.

With regard to "GaH_3" sources, as applied to MOCVD, $GaH_3 \cdot AsH_3$ would appear to have potential - as yet this species has not been reported. Furthermore, based on the decreased Lewis base strength, as one proceeds down the group VI hydrides, one would not expect $GaH_3 \cdot AsH_3$ to be stable at 25°C in the absence of additional bonding interactions.

Another hydride source for gallium is the previously mentioned $Ga_2Me_4H_2$[20], however, this species is reported to decompose at 130°C; furthermore it obviously contains carbon, a potential impurity concerning MOCVD. Gallium borohydride, $Ga(BH_4)_3$, is yet another volatile compound of gallium, however, it is very unstable decomposing well below room temperature.[21] The related species, $(CH_3)_2GaBH_4$ was first prepared via the interactions described by eq. 7,[21] however,

$$Ga(CH_3)_3 + B_2H_6 \longrightarrow (CH_3)_2GaBH_4 + 1/2(MeBH_2)_2 \qquad (7)$$

more recently Downs[22] has described the preparation according to eq.8

$$(CH_3)_2GaCl + LiBH_4 \longrightarrow (CH_3)_2GaBH_4 + LiCl \qquad (8)$$

Again, $(CH_3)_2GaBH_4$ is thermally unstable at 25° as described eq. 9.

$$2(CH_3)_2GaBH_4 \longrightarrow 2Ga + 3H_2 + BMe_3 + 1/2[CH_3BH_2]_2 \qquad (9)$$

A final non-carbon containing volatile gallium compound has been reported, $HGa(BH_4)_2$, however, this species is also unstable well below room temperature.[23]

B. Fluorocarbon Derivatives of Gallium.

The increased C-F bond strength over that of C-H, 116vs.98 k cal/mole, contributes to the greater thermal stability of the perfluorocarbon moriety over the analogous hydrocarbons fragment. Furthermore, decreased intermolecular forces between fluorocarbon metal derivatives may result in increased volatility over the analogous hydrocarbon species. To data, perfluorocarbon derivatives of gallium are unknown; however, recent advances in synthetic techniques, may well afford $Ga(CF_3)_3$.[24] With regard to species such as $Ga(CF_3)_3$, the increased (CF) bond strength, over that of $Ga(CH_3)_3$ (CH) may well produce GaAs containing a very low carbon impurity concentration.

Two methods are suggested as possible routes to the formation of $Ga(CF_3)_3$, (1) exchange of GaX_3 (where X is a halide) with $Hg(CF_3)_2$, and (2) generation of CF_3 radicals by discharge reactions, and subsequent radical reactions with gallium sources.

Although $Hg(CF_3)_2$ had previously been demonstrated not to undergo ligand exchange with other metals[25] (unlike $Hg(CH_3)_2$), recent reinvestigations of this type of reaction have demonstrated its feasibility, and the importance of controlling the reaction parameters. For example, treatment of GeI_4 with $(CF_3)_2Hg$, while controlling the reaction temperature, time of reaction, and mole ratio, affords $Ge(CF_3)_4$.[26] With regard to RF discharge reactions, this technique results in low thermal reaction conditions thereby allowing the isolation of species of relatively low thermal stability. Using this technique HgI_2 and $GeBr_4$, separately, were subjected to CF_3 radical to afford $Hg(CF_3)_2$ and $Ge(CF_3)_4$ respectively.[24]

Another derivative of gallium which may have potential as a pure source for gallium, as related to MOCVD, is $Ga(CF_2CF_2H)_3$, as yet unknown. This

species is envisioned to be a volatile source of insitu gallane, GaH_3, via β elimination eq.10, and the by-product, perfluoroolefin, would not be expected to readily decompose to afford carbon thereby contaminating the GaAs formed.

$$Ga(CF_2CF_2H)_3 \longrightarrow "GaH_3" + 3CF_2{=}CF_2 \qquad (10)$$

V. RECOMMENDATIONS:

When one compares the epitaxial GaAs produced from trialkylgallium species and arsine, the triethyl rather than trimethyl derivatives apparently afford a product of higher quality. This is particularly true for sub-atmospheric reaction conditions and this result is herein interpreted in terms of a β elimination reaction mechanism. It is strongly recommended that future MOCVD reactor designs encorporate the provision for low pressure reactions, i.e. to 10 torr. When comparing sources of (GaR_3, R=CH_3 and C_2H_5), one must also consider the purity of GaR_3 in evaluating which species is "best" for a particular MOCVD reactor design. Of course the purity of arsine must also be taken into consideration along with that of any carrier gases.

The chemistry of a number of volatile gallium compounds has been discussed as related to new gallium sources for MOCVD. In particular $GaH_3{\cdot}N(CH_3)_3$ has been mentioned, of course carbon incorporation may occur with this species. Although one would expect $N(CH_3)_3$ to exhibit stronger nucleophilic character than NH_3, the application of NH_4Cl, eq. 4, is recommended in an attempt to produce $GaH_3{\cdot}NH_3$. The latter then could be utilized in conjunction with AsH_3 to produce GaAs. Incorporation of N in GaAs should not detract from this chemical system because of the non-overlap of band gaps for N and GaAs. In a related fashion, the synthesis and application of $GaH_3{\cdot}AsH_3$ toward GaAs formation should be investigated.

A potential source for low carbon impurity GaAs involves the application of per and polyfluorocarbon derivatives of gallium. It is strongly recommended that the preparation of $Ga(CH_2CF_2H)_3$ be pursued and that, if successful, its mode of thermolysis be investigated as related to insitu formation of "GaH_3". In a related manner, percursors to the formation of $Ga(CF_2CF_2H)_3$, may result, eg. $H_{3-x}Ga(CF_2CF_2H)_x$, where x=1 or 2. These latter species may be stabilized by adduct formation with AsH_3 due to the increased electronegativity of Ga as a result of polyfluoroalkyl substitution for H.

REFERENCES

1. H. M. Manasevit, Appl. Phys. Letters 12 (1968) 156.

2. H. M. Manasevit and W. I. Simpson, J. Electrochem. Soc., 116 (1969) 1725.

3. P. D. Dapkus, H. M. Manasevit, K. L. Hess, T. S. Low, and G. E. Stillman, J. Crystal Growth, 55, (1981) 10.

4. S. J. Bass, J. Crystal Growth, 31, (1975) 172.

5. J. P. Duchemin, M. Bonnet, F. Koelsch, and D. Huyghe, J. Crystal Growth, 45 (1978) 181.

6. S. Ito, T. Shinobara, and Y. Seki, J. Electrochem. Soc., 120 (1973) 1419.

7. C. Y. Change, Y. K. Su, M. K. Lee, L. G. Chen, and M. P. Houng, J. Crystal Growth, 55, (1981) 24.

8. Y. Seki and K. Lanno, K. Iida and E. Ichiki, J. Electrochem. Soc., 122 (1975) 1108.

9. Y. Nakayama, S. Ohkawa, H. Hashimsto, and H. Ishikawa, J. Electrochem. Soc., 123 (1976) 1227

10. K. Lindeke, W. Sack, and J. J. Nickl, J. Electrochem. Soc., 117 (1970) 1316.

11. A. Zaouk, A. Lebugle, and G. Constant, J. Crystal Growth, 46, (1979) 415.

12. A. Zaouk, E. Solvetat, J. Sakaya, F. Maury and G. Constant, J. Crystal Growth, 55, (1981) 135.

13. K. Ziegler, K. Nagel, and W. Pfohl, Annalen, 629, (1960) 210.

14. G. E. Coats, "Organometallic Compounds", 2nd ed., p. 144 Methuen, London, 1960.

15. N. N. Greenwood, A. Storr, and M. G. H. Wallbridge, Proc. Chem. Soc., p. 430 (1962).

16. D. F. Shriver, R. L. Amster, and R. C. Taylor, J. Am. Chem. Soc., 84, (1962) 1321.

17. A. F. Finholt, A. C. Bond, Jr., and H. I. Schlesinger, J. Am. Chem. Soc., 69, (1947) 1199.

18. E. Wiberg, and M. Schmidt, Z. Naturforsch., 66, (1951) 172.

19. D. F. Shriver, R. W. Parry, N. N. Greenwood, A. Storr, and M. G. H. Wallbridge, Inorg. Chem., 2, (1963) 867.

20. E. Wiberg and T. Johannsen, Angew. Chemie., 55, (1942) 38.

21. H. J. Schlesinger, H. C. Brown, and G. W. Schaeffer, J. Am. Chem. Soc., 65 (1943) 1786.

22. A. J. Downs, P. D. P. Thomas, J. Chem. Soc., Dalton, (1978) 809.

23. A. J. Downs and P. D. P. Thomas, Chem. Comm., (1976) 855.

24. R. J. Lagow and J. A. Morrison, "Advances In Inorganic and Radiochemistry", 23, (1980) 177. Academic Press., Inc.

25. H. C. Clark, "Adv. Fluorine Chem.", 3, (1963) 19.

26. J. A. Morrison, L. L. Gerchman, R. Euken, and R. J. Lagow, J. Fluorine Chem., 10 (1977) 333.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

SOFTWARE SIMULATION REQUIREMENTS FOR THE CVA PROGRAM AND OTHER APPLICATIONS

Prepared by: Dr. K. Sam Shanmugan
Academic Rank: Professor
Department: Electrical Engineering Department
University: University of Kansas
Research Location: RADC, Griffiss AFB, Rome NY
USAF Research Colleague: Mr. Peter Leong
Date: July 30, 1982
Contract No: F49620-82-C-0035

SOFTWARE SIMULATION REQUIREMENTS FOR THE CVA PROGRAM AND OTHER APPLICATIONS

K. Sam Shanmugan

ABSTRACT

The DVAL joint test force methodology for CVA requires a considerable amount of pre-test and post-test analysis of receiver performance under a variety of jamming environments. The complexity of the receivers and the jamming environments preclude the use of a purely mathematical approach for receiver performance analysis. A combination of mathematical methods along with Monte-carlo type simulation of the receiver and the link will lead to a fast and accurate method of pre test and post-test analysis. This report outlines a simulation approach and the of simulation software requirements for the CVA program.

While the CVA program is used as one example where simulation can play a significant role, a general purpose communication systems simulation software package can be used in a variety of applications within RADC.

ACKNOWLEDGEMENT

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to spend a very worthwhile and interesting summer at the Rome Air Development Center, Griffiss AFB NY. The hospitality of the DCLF Section and the Communications Branch is greatfully acknowledged.

The author would also like to express his thanks to Mr. Peter Leong, and Major Lee Wagenhals for their collobaration and guidance during the course of the summer research program.

TABLE OF CONTENTS

1. Introduction

2. Objectives

3. Simulation Support for CVA Analysis

4. Over-all Requirements for a Communication Link Simulator

    3.1 Model Library
    3.2 Flexible Topology
    3.3 Parameter Modification
    3.4 Speed and Accuracy
    3.5 Test Point Selection and Post-Processing
    3.6 Transportability
    3.7 Documentation
    3.8 Comparison of Some of the Existing Packages

5. Library of Models

    5.1 Sources
    5.2 Signal Conditioners, Source Encoders and Encryptors
    5.3 Error Control Coders and Decoders
    5.4 Modulators and Demodulators
    5.5 Channels
    5.6 Filters
    5.7 Synchronizers
    5.8 Antennas
    5.9 Miscellaneous
    5.10 Estimators

6. Summary and Recommendations

7. References

1. INTRODUCTION

Assessment of the vulnerability of a radio frequency data link subjected to intentional electromagnetic interference is an extremely difficult and complex task. In a recent report (1) the DVAL joint test force had outlined a four phase step-by-step procedure for determining the anti-jamming capabilities of RF links. The four major phases of the DVAL methodology address the following areas:

1) Susceptibility of the receiver
2) Interceptibility of the transmission
3) Accessibility of the link/receiver to the jammer
4) Feasibility of degrading receiver performance.

Each of the four phases are divided into three major steps:

1) Pre-test analysis
2) Test design and execution
3) Post-test analysis

A combination of analysis, measurements, and database interaction provides the basis of the DVAL methodology.

Of the three steps in each phase of the DVAL methodology, the test design and execution step covers activities that could be potentially very expensive and time consuming. It is therefore necessary to reduce to the absolute minimum those issues which need to be resolved in the test-related steps of the CVA. This can be done by attempting to resolve as many issues as possible during the pre-test analysis phase. Indeed, the overall success of the CVA program might depend, to a large degree, on the speed and accuracy with which the pre-test analysis can be carried out.

The analysis of the susceptibility, interceptibility, and accessibility of a data link in a jamming environment can be carried out using either mathematical approaches alone or computer simulations alone or a combination of the two. The complexity of modern communication links along with the presence of several nonlinear elements (limiters, envelope detectors, high power amplifiers) preclude the use of a totally

analytical approach. The assumptions and simplifications needed to obtain closed form analytical results are often very restrictive, and the final mathematical model is often not a very good representation of the system. It may sometimes be possible to derive performance bounds (upper and lower bounds) using analytical methods without too many restrictive assumptions, but the bounds may not be tight enough to carry out detailed analysis of the cumulative effects of several degradation mechanisms in a communications link.

In recent years, computer-aided modeling and simulation of communication links has proven to be a valuable addition to analytical and hardware measurement approaches to predicting performance (2-13). With the simulation approach, sampled signals are created in the computer and then operated by algorithms that simulate the effects of coding, filtering, detection, synchronizing, and other operations that take place in a communication link. As the simulation progresses, waveforms at preselected data extraction points (DEPS) are stored for later analysis to determine measures of performance (MOPS).

Monte-Carlo type simulations using detailed models of functional blocks connected in a free topological configuration provide a flexible and accurate method for analyzing communication links and networks. Some of the software packages combine simulation and analytical approaches to improve the speed and accuracy. Many software simulation packages are now used extensively as CAD tools for communication systems design and the validity of the results have been well documented (2-12).

The pre-test and post-test analysis requirements of the DVAL JTF methodology for CVA can best be met by a combined mathematical and simulation approach. This report recommends such an approach and defines the requirements of a software simulation package for supporting the CVA program.

Besides supporting the CVA program, the software simulation package can be used for a number of other applications. For example, simulations can be used to estimate the performance of new and innovative systems that are being proposed by RADC engineers as well as outside contractors. Detailed design trade-off studies, and comparison of competing design ideas can also be carried out using simulations. By encouraging all the contractors as well as RADC engineers to make use of the simulation package to validate their design claims, an objective basis for comparison can be established. A facility for performing such trade-off studies and for analyzing a wide variety of communication links does not exist within the DOD at the present time. While RADC/DICEF has been working on developing such a facility, this development effort is now at a critical point in that a considerable amount of software development is required before the full capabilities of the software simulation packages already developed in parts by DICEF can be demonstrated and utilized. The main objective of the summer research program was to define the software development needed for supporting the CVA program and other applications.

2. OBJECTIVES

Specific objectives of the summer research program were:

1. To evaluate the simulation requirements for the CVA program and other applications,

2. to perform a comparative evaluation of the software simulation package ICSSM, and

3. to recommend a plan for developing software simulation models to support the CVA program and other applications.

Section 3 of this report presents a simulation approach to DVAL pre-test and post-test analysis, in particular the pre-test analysis. The overall requirements of a software simulation package are given in Section 4, and a list of functional block models that are needed is given in Section 5. Specific recommendations for

developing the software simulation capabilities to support the requirements of the CVA program and other applications within DC are presented in Section 6 of the report.

3. SIMULATION SUPPORT FOR CVA ANALYSIS

The vulnerability of a communication system can be assessed at many different levels starting from small functional blocks in a receiver to the network level. Irrespective of the level at which the assessment is made, the vulnerability issues have to be resolved based on analysis and test results. A major component in a vulnerability assessment program is the analysis of the susceptibility and accessability of the communication link to a jammer. This analysis is often mathematically untrackable bacause of the presense of nonlinearities, heuristic decision algorithms, and complexity. In recent years, computer simulations have been used successfully as a viable alternative to purely mathematical analysis, and the simulation approach can play an important role in the analysis phases of the CVA program.

As an example, the simulation approach is almost tailor made for the pre-test analysis phase of assessing the susceptibility of an RF link. The pre-test analysis tasks shown in Figures 3.1 and 3.2 can be easily carried out using a simulation package that has the following features:

a) A large library of mathematical models of functional blocks in a communication system.

b) A high level language or interactive input for selecting and connecting functional blocks with appropriate parameters into a receiver structure that is similar to the actual receiver implementation.

c) A simulation exercisor that executes the simulation model and stores the waveforms at selected data extraction points (DEPS).

d) A post-processor that can analyze the stored waveforms at selected DEPS and calculates the measures of performance (MOPS).

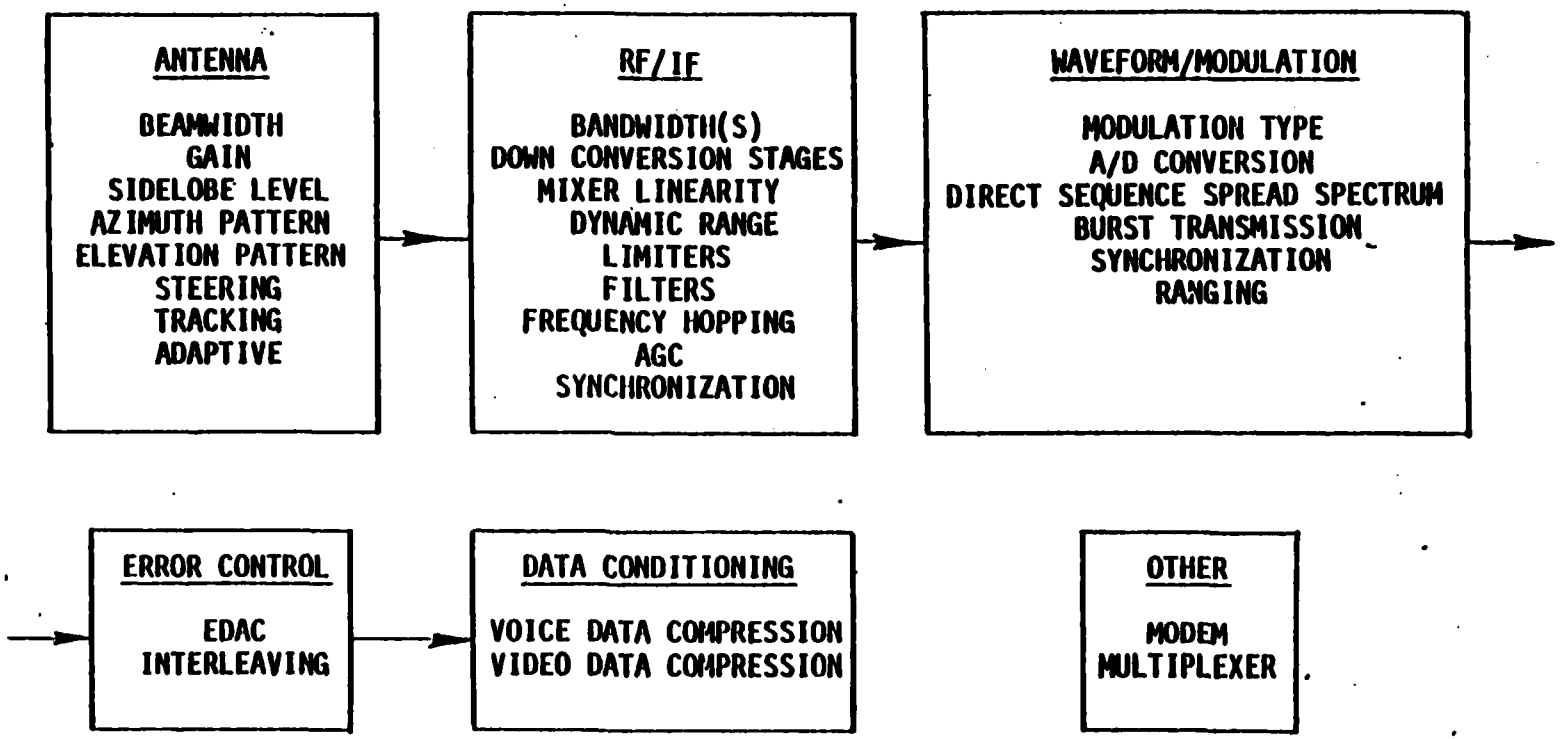


Figure 3.1. The Pretest Analysis Task Flow

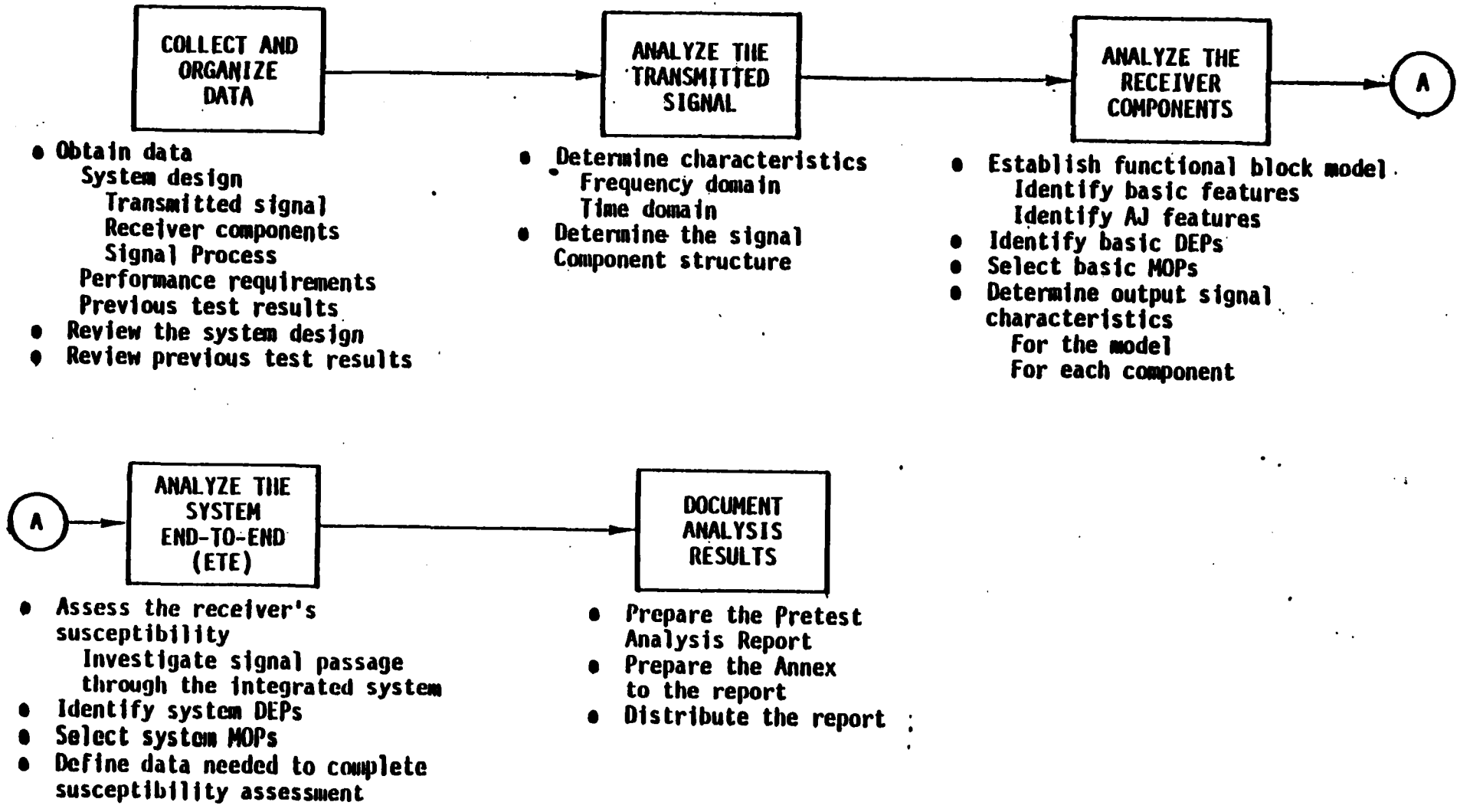


Figure 3.2. Receiver Characteristics by Functional Block

Given a software simulation package with the capabilities listed above, then the analyst/engineer can start from a block diagram of the system to be simulated, select the functional blocks from the library, specify the interconnections and parameter values, execute the model and the post-processor, and obtain answers to the vulnerability issues being addressed. The analysist can address and obtain answers to vulnerability issues for a functional block or a subsystem or the receiver or for the communication link as a whole. With a good simulation package the analyst can duplicate the actions of a well-trained design and test engineer without the high cost of hardware design and testing. The simulation package can also be used to address vulnerability issues during the design phases of receivers and other components.

The simulation technique that is recommended for supporting the CVA program is one in which the signals and signal processing operations mimick the actual waveforms and hardware operations. The mathematical simulation models of hardware functional blocks and subsystems should be capable of handling any level of detail specified by the analysist. At the same time, the simulation package should provide simple models for large subsystems if the analyst is interested in a large scale system of which a particular subsystem is a small part that has been analyzed in detail before. As an example, suppose the analyst is interested in evaluating the performance of a convolutional error correction code. Then, the simulation package should provide detailed models for the components of the encoder and decoder, and use a simple finite state discrete channel model for representing the physical channel and all the components of the transmitter and the receiver. Of course the parameters of the channel model has to be obtained from a detailed simulation of the physical channel, the transmitter and the receiver. This hierarchical structure in which the higher level componnents have a simpler representation compared to the detailed models of the lower level components is essential for the simulation of very complex communication systems.

With a properly structured simulation package, vulnerability issues can be addressed at various levels - at the level of functional blocks, receivers, physical links, communication links, and at the network level. At the present time the vulnerability issues addressed in the DVAL JTF methodology refer mostly to the

physical link (transmission media) and the receiver. At this level, a well structured communication link simulation package can provide answers to the following questions.

1. How do data link system errors characteristically relate to interference power and type of modulation?

2. What interference power levels are required to significantly degrade the data link for selected types of interference modulation?

3. What are the predominant degradation mechanisms for each selected interference waveform?

4. What are the optimum modulation parameters for interference? (Is there an interference waveform which is significantly more effective than other forms of modulation?)

5. What is the data link receiver response to combinations of interference modulations?

6. How effective are the AJ features of the data link receiver?

It should be emphasized that many network level vulnerability issues have to be addressed and procedures for assessing network susceptibility have to be developed in the near future. The communication link simulation software should be developed in such a fashion that it can interface easily with a network simulation package.

4. OVERALL REQUIREMENTS FOR A COMMUNICATION LINK SIMULATOR

As mentioned in an earlier section of this report, the vulnerability of a communication system can be assessed at different levels and hence the analysis can also be carried out in a hierarchical fashion with different levels of detail as illustrated in Figure 4.1. While simulation software for computer aided analysis at the three levels shown in Figure 4.1 is necessary for assessing issues related to the vulnerability of the network as a whole, the development and use of the software

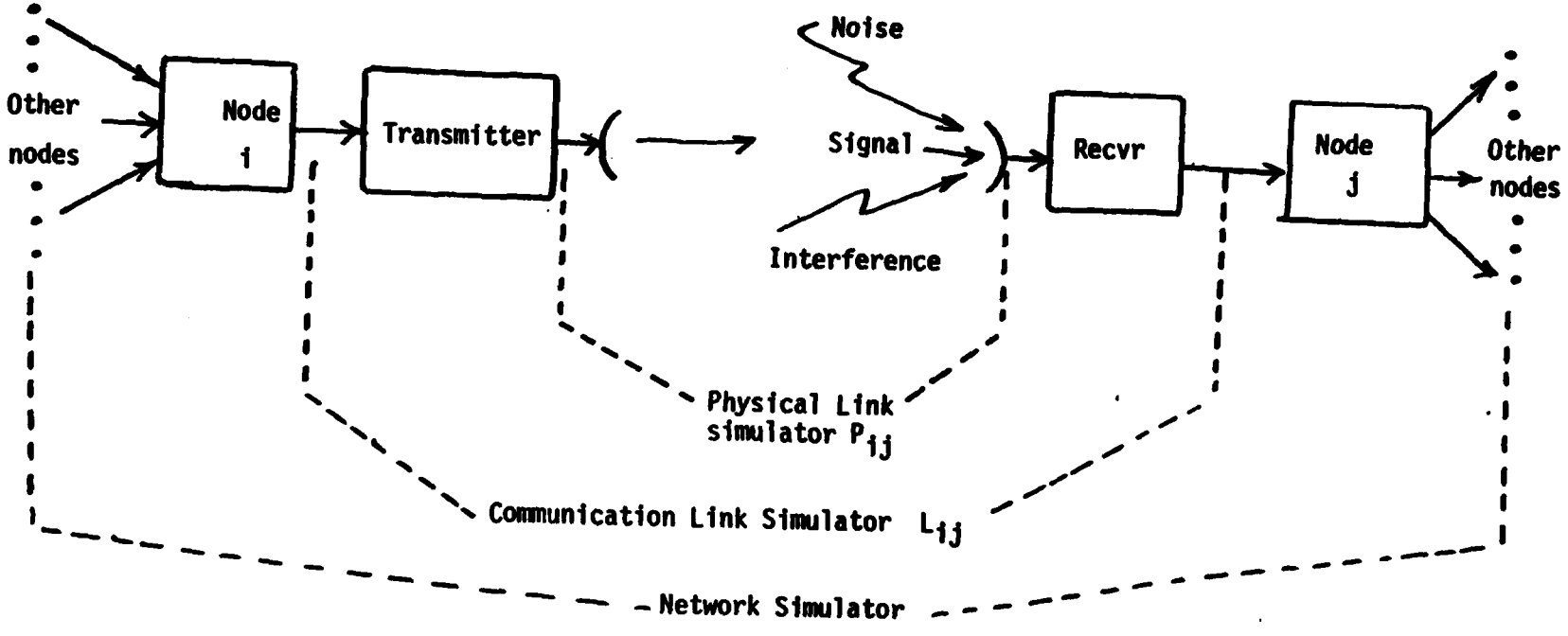
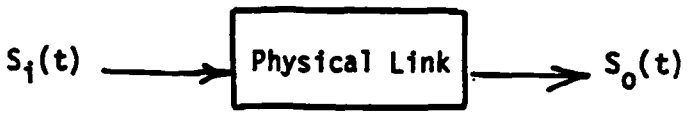


Figure 4.1. Hierarchical Simulation Structure



$$S_o(t) = \underbrace{S_i(t)a(t)}_{\text{Signal}} + \underbrace{\sum_{k=1}^{n} S_i(t-t_k)d_k(t)}_{\text{Multi-path}} + \underbrace{\sum_{k=1}^{m} J_k(t)a_k(t)}_{\text{Interference}} + \underbrace{n(t)}_{\text{Thermal noise}}$$

$S_i(t)$ = Complex valued input signal; $S_o(t)$ = Complex valued output signal

$a, a_k(t)$ = time-varying adaptive antennae gain

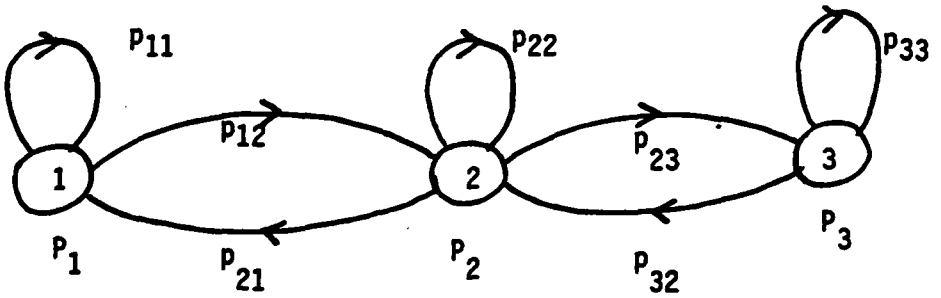
$j_k(t)$ = jamming signal

$d_k(t)$ = time-varying attenuation for multi-path

$t_k(t)$ = delay of the k-th multi-path

$n(t)$ = thermal noise

Figure 4.2.a. Example of a Physical Link Model to be Used in a Link Simulation



State 1 - random errros, error probability = PE_1
State 2 - transition state, error probability = PE_2
State 3 - bursty error, error Probability = PE_3

P_i = Probability of being in state i, p_{ij} = transition probability

R_c = state transition rate

Figure 4.2.b. Example of a three state communication channel model for use in network simulation

can be carried out in parallel fashion. As illustrated in Figure 4.2, for example, the simulation of a communication link can be done with great attention to the details of the transmitter and receiver while treating the physical link or the transmission media as a single block that is described in a generic fashion (i.e. say as a fading dispersive channel with interference and Gaussian noise). Thus issues related to the susceptibility of the receiver can be addressed with detailed models while the physical channel is treated as a black box. In a similar vain, issues such as the susceptibility of the network synchronizing, protocols and control can be addressed with simpler models for the communication links.

While the hierarchical approach outlined above may have some limitations, it does provide a framework for starting the development of software support for the CVA program. The requirements for a software package for simulating a communication link are addressed in this report under the assumption that packages such as TACOM already exist for the physical link simulation and that a network simulation package will be acquired in the future or developed in parallel with the communication link simulation software development.

The major components of a software package for simulating communication links are the following:

1) A large library of mathematical models for hardware functional blocks in a communication system.

2) A model configurator that accepts commands in a user oriented language for selecting and connecting a set of functional blocks with appropriate parameters.

3) A simulation exercisor that executes the simulation model and produces fast and accurate results representing the waveforms at selected data extraction points, and

4) a post-processor that can analyze the stored waveforms, compare measures of performance and display the result.

The following sections of the report address the overall requirements of the major components of a software package for analyzing communication links. A brief comparison of the capabilities of some of the existing software packages is also presented.

4.1 MODEL LIBRARY: In order to simulate a variety of existing and proposed communication systems, the simulation package should contain an extensive library of pre-programmed functional block models. The library should contain a variety of signal generators, signal conditioners, source encoders/decoders, error control encoders/decoders, modulators, demodulators, encryptors/decryptors, synchronizers (acquisition and tracking), filters, waveform analyzers, and routines for calculating performance measures. The software package should have provisions for entering into the library user developed models in the form of self contained FORTRAN subroutines as well as subsystem models previously configured using previously developed functional models.

4.2 FLEXIBLE TOPOLOGY: The software simulation package (SSP) should allow the user analyst to specify any topology that he desires for his system. However, the SSP should check the topology for improper or missing interconnections and provide diagnostic messages to aid the user analyst in correcting any errors. This is best accomplished by using a menu-driven interactive process for configuring the system model.

4.3 PARAMETER MODIFICATION: Provisions should exist for (easily) specifying a set of initial parameters and running several iterations by changing one or more parameter values.

4.4 SPEED AND ACCURACY: Monte-Carlo simulations often require the processing of millions of samples and the accuracy of direct Monte-Carlo simulations are inversly related to the square root of the sample size. The SSP should have provisions for improving the speed and accuracy of simulations i.e. provide for better accuracy (smaller simulation induced vairance) and smaller sample size requirements. This can be accomplished by incorporating the following features:

a) Block processing via FFT (not appropriate when feedback is present)

b) Model partitioning in which different numbers of samples are processed in different modules.

c) Incorporating and using analytical results in calculating measures of performance such as probability of error.

d) Using extrapolative and biased sampling schemes which reduce the sample size requirements considerably (13)

4.5 TEST POINT SELECTION AND POST PROCESSING: The SSP should make provisions to allow the user analyst to select the data extraction points and allow for the post processing of the waveforms at the DEPS. The post-processing routines should simulate a variety of measurement devices that the analyst might want to use to obtain measures of performance. Typical post processing analysis routines that should be included in the SSP are spectrum analyzers, power meters, bit error estimators and statistical analysis routines to compute and plot histograms etc. The test point selection and post-processing should allow the user analyst to simulate not only actual hardware tests but also allow him to observe and process waveforms without regard to such provisions in the communication link hardware. The post-processing routines should have a full complement of well structured plotting/display routines.

4.6 TRANSPORTABILITY: The SSP should be written in a higher level language such as FORTRAN, PASCAL or PL/1 so that it can be transported between installations. Of course, some conversion effort will be needed to accommodate differences in operating system characteristics in I/O and file handling.

4.7 DOCUMENTATION: This is one of the most important requirements from a user point of view. The extent to which a SSP will be used will depend on how good the software documentation is. In addition to an easy to follow users manual, the documentation should include help files within the software package that the user can call while he is using the various components of the package. Besides good

software documentation, the algorithms used in various subroutines should be explained in the tutorial fashion in a separate manual.

3.8 COMPARISON OF SOME OF THE EXISTING SOFTWARE SIMULATION PACKAGES:

A number of software simulation packages, some of which are commercially available, meet the requirements set forth in the preceeding sections to some extent. While it is not possible to compile a listing of the features of all the packages that are available at the present time, there are three candidate packages whose features are summarized in Table 4.1.

The Interactive Communication Systems Simulation (ICS) package was developed for DICEF by RPI and it runs on a dedicated minicomputer with an attached array processor (10). The most attractive features of ICS are its speed of execution, good user interface and graphics capabilities, and an adequate model library. The most serious limitations of the ICS package are its use of fixed topology and its dependence on a particular host machine and array processor (of course these are the features that contribute to one of the main strengths of ICS, namely its speed).

Another simulation package, Interactive Communication Systems Simulation Model (ICSSM), developed by the Hazeltine Corporation for DICEF, runs under the MULTICS operating system at RADC (11). ICSSM allows for flexible topology specified by the user via an interactive, menu-driven model configurator. The main strengths of ICCSM are its interactive model configurator that has many attractive features, and its library sructure which has provisions for easily adding user derived models. The main limitations of the ICCSM package are that its model library is almost empty, and that it requires a considerable amount of bookkeeping on the part of the user in the model development phases. With more experience, this overhead can be considerably reduced. As more users become familiar with the model structure, the sharing of expertise will greatly contribute towards easing the model interfacing requirements. Before ICSSM can be used to address CVA issues, the model library has to be enhanced considerably.

SYSTID is a simulation package that was developed by the Hughes Aircraft Company and is available commercially on a limited basis. This software package

had been used by a number of organizations during the past 10 years, and the results are well documented in the open literature (5, 8, 12). The main strengths of SYSTID are its easy to use input language structure, free topology, a large collection of validated library models, and an excellent post-processor package. Major deficiencies of SYSTID include insufficient software documentation, and lack of models for adaptive antenna arrays and spread spectrum systems. While models for these systems can be constructed (though not easily) using existing SYSTID models, no generic models exist for the major components of adaptive arrays and spread spectrum systems. Such models do not exist in ICS or ICSSM either.

Either SYSTID or ICSSM can serve as the framework for providing simulation support for the CVA program. The availability of SYSTID is limited and a considerable amount of conversion effort will be required if SYSTID is to be implemented at RADC. Since ICSSM is already implemented and tested at RADC, and since ICSSM has some additional capabilities compared to SYSTID in the model configuration area, it is recommended that ICCSM be the simulation system for providing the required simulation support for the CVA program.

ICS can play an important supportive role in the communication systems simulation efforts at RADC. Some applications such as the evaluation of error control coding schemes can be simulated with excellent speed by using ICS. ICS can also be used to cross validate some of the models that will have to be developed for ICSSM. Indeed, many of the algorithms used in ICS routines can and should be adopted for implementation as ISSCM modules.

FEATURE	ICS	ICCSM	SYSTID
1) Model library			
a) overall	good	ND	very good
b) generic models for adaptive arrays	UD	ND	fair
c) generic models for spread spectrum	UD	ND	fair
2) Topology specification	fixed	free	free
3) User interface	very good	very good	very good
4) Speed			
a) overall	excellent	good	good
b) partitioning/restart	no	yes	yes
c) modified Monte-carlo/ other techniques	no	no	fair
5) Post-processing/graphics	good	ND	Excellent
6) Transportability	poor	good	good
7) Users manual	good	good	fair
8) Software Documentation	fair	good	fair
9) Intermixing of model configuration statements and FORTRAN	no	no	yes

ND-not developed; UD-under development

Table 4.1 Comparison of three simulation packages

5. LIBRARY OF MODELS

In order to serve the needs of a wide variety of users, the software simulation package should have a sufficiently large number of preprogrammed simulation models of functional building blocks to accommodate most communication link configurations. While it is not possible to fully anticipate the needs of all potential users, having models for a large set of typical functional blocks encountered in communication systems will minimize the need for user developed programs. The SSP should have provisions for adding user developed FORTRAN models to the library if a library model is not available for a particular component that a user is interested in. As the number of users grow, the library will also grow to meet the needs of the users.

At the present time, ICSSM has very little to offer in the way of library models. A list of models that should be added to the ICSSM library is given below. A brief description of how these models can be used to simulate a particular spread spectrum modem is given in Appendix A of the report.

The library models are grouped under the following categories:

1. Sources
2. Signal Conditioners, Source Encoders and Encrypters
3. Error Control Coders/Decoders
4. Modulators and Demodulators
5. Channels
6. Filters
7. Synchronizers
8. Antennas
9. Miscellaneous
10. Estimators (MOPS)

5.1 SOURCES

1.a. Signals

1. Periodic Waveforms
2. Pulse and Impulse
3. Damped Oscillations
4. Raised Cosine Shape Generator
5. Random Binany Waveform
6. M-ary Random PAM Waveform
7. PN Sequence Generator (LFSR Type)
8. Bi-phase/Manchester PN Generator

1.b. Noise

1. White Gaussian
2. Arbitrary psd (non Gausian noise source)
3. Impulse Noise

1.c. Interference Waveforms

1. CW Tone
2. CW Tone Bursts
3. Noise Loaded AM
4. Noise Loaded FM
5. Direct Noise
6. Synchronous

Some of the library models will be completely self contained, while other models might consist of a collection of previously developed models. For example a low pass filter model, and a Gaussian noise source may be self contained modules, where as a model for additive dispersive white Gaussian noise channel (AWGN) will make use of the low pass filter model and the Gaussian noise source model.

5.1 SOURCES

1.a. Signals

1. Periodic Waveforms
2. Pulse and Impulse
3. Damped Oscillations
4. Raised Cosine Shape Generator
5. Random Binany Waveform
6. M-ary Random PAM Waveform
7. PN Sequence Generator (LFSR Type)
8. Bi-phase/Manchester PN Generator

1.b. Noise

1. White Gaussian
2. Arbitrary psd (non Gausian noise source)
3. Impulse Noise

1.c. Interference Waveforms

1. CW Tone
2. CW Tone Bursts
3. Noise Loaded AM
4. Noise Loaded FM
5. Direct Noise
6. Synchronous

5.2. SIGNAL CONDITIONERS, SOURCE ENCODERS & ENCRIPTORS

1. A/D Converters
2. D/A Converters
3. Multilevel PCM
4. Transform Encoders
5. Multiplexers/Demultiplexers
6. Concentrators/Distributors
7. Burst Conditioner/Generator
8. Table Look Up Type Encryptor/Decryptor
9. DES Type Encryptor/Decryptor

5.3. ERROR CONTROL CODERS/DECODERS

1. Block Encoder (Hamming, Golay, BCH)
2. Interleaver/De-interleaver
3. Convolutional Encoder
4. Table Look Up Decoder for Block Codes
5. Hard Decision Decoder for Block Codes (for Hamming, Golay, BCH codes)
6. Viterbi Decoder for Convolutional Codes
7. Soft Decision Symbol by Symbol Decoder for Block Codes

5.4. MODULATORS AND DEMODULATORS

a. Amplitude Modems

1. DSB-SC Modulator
2. AM Modulator
3. SSB Modulator
4. VSB and VSB + C
5. Envelope Detector
6. Coherent Detector

b. Angle Modems

1. Phase Modulators
2. Frequency Modulator
3. Discriminator
4. PLL Demodulator
5. Zero-crossing Detector

c. Baseband Pulse Modulation

1. PCM-Binary
2. Differential PCM
3. M-ary PCM/PAM
4. Delta Modulator
5. CVSD Modulator
6. Pulse Position Modulator
7. Integrate and Dump Demodulator
8. Sample/Hold D to A

d. Digital Carrier Modulation

1. ASK
2. PSK
3. FSK
4. QPSK/OQPSK
5. MSK
6. QAM

5.5. CHANNELS

1. Additive White Gaussian Noise
2. Additive Impulse Noise
3. Additive Interference
4. Multiplicative Noise
5. Fading/Dispersion/Multipath
6. BSC
7. M-ary Discrete
8. M-ary Discrete with Memory

5.6. FILTERS

1. Ideal Lowpass, Highpass and Bandpass
2. Butterworth
3. Chebyshev
4. Elliptic
5. Arbitary Poles and Zeroes
6. Specified Amplitude and Phase Response
7. Matched Filter
8. Tapped Delay Line Filter
9. Adaptive Transversal Equalizer

5.7. SYNCHRONIZERS

1. Tapped-delay Line Correlator
2. Bit Synchronizer
3. Phase Tracking Loop (open and closed)
4. Direct Sequence Aquisition-sliding Correlator
5. FH-Coarse Aquisition
6. Delay-locked Loop Tracking
7. Dither Loop Tracking

5.8. ANTENNAS

1. Omnidirectional
2. Directional
3. Adaptive
   - LMS Algorithms
   - Mathematical Tracking Model

5.9. MISCELLANEOUS

1. Delay Lines (fixed & random delays)
2. Limiters
3. Differentiator
4. Integrator
5. Summer
6. Multiplier
7. Comparator
8. Phase Shifter
9. AM-AM Nonlinearity
10. AM-PM Nonlinearity
11. Buffers

5.10. ESTIMATORS OF MEASURES OF PERFORMANCE (POST-PROCESSORS)

1. Power Meter
2. Noise Bandwidth Meter
3. Time Delay Estimator
4. Spectrum Analyzer
5. Eye Pattern Generator
6. Signal Constellation Plotter
7. Signal Locus Plotter
8. Bit Error Estimator
   a. Error Probability
   b. Burst Analysis
9. Statistical Analyzer (mean, variance, histogram etc)

All of the models listed above need not be developed in a single effort. The models in the list can be prioritized with a focus on a set of initial applications that are of interest. The remaining models can be developed as the need arises.

6. SUMMARY AND RECOMMENDATIONS

Sofware simulation can play a significant role in the analysis and design of communication systems. The CVA program and other applications can make use of simulations to obtain answers to performance related questions for which purely mathematical approaches do not provide satisfactory answers in a timely fashion. Simulations can also be used to reduce the amount of costly hardware tests necessary for assessing system performance. During the design phases, software simulations can be used to carry out design trade-off studies.

The software simulation package ICSSM, developed for RADC by the Hazeltine Corporation, provides an excellent frame work within which communication link simulations can be carried out. To demonstrate and utilize the full capabilities of ICSSM, an effort should be undertaken to develop an extensive set of subroutines or software models of hardware functional blocks encountered in a wide variety of communication systems. It is recommended that this effort be carried out in two concurrent phases:

1. A short term software model development for two immediate applications (SEEK COMM modem vulnerability assessment, and evaluation or proof of concept for a new modem)

2. A long term software model development effort to build up an extensive set of library models.

During the two phases of the software development, particular attention should be paid to (1) good documentation of programs and algorithmns (2) improved simulation speed via modified Monte-Carlo and other techniques, and (3) model validation via some hardware measurements and cross validation with results obtained using other simulation packages such as ICS or SYSTID.

With successful completion of the two phases, RADC will have a unique capability to simulate a variety of communication links. Without the additional software development effort recommended in this effort, the currently available software will be of very limited use and it is unreasonable to expect the project engineers at RADC to develop the necessary software models in house. If the proposed software development could not be undertaken and completed in a timely fashion, then RADC and DICEF might as well give up the communication link simulation effort.

Another area that deserves careful study at this point in time, is communication network simulation. Several network simulation packages are commercially available. A study can be initiated to comparatively evaluate the capabilities of these packages and one of the packages can be aquired and interfaced to the link simulation software. This will give RADC the added capabilities to perform communication systems simulation at various levels in a hierarchical fashion.

7. REFERENCES

1. Data Link Vulnerability (DVAL) T&E Methodology, volumes 1-4, DVAL Joint Test Force, Kirtland AFB, New Mexico, 87117, 1981

2. D.A. Lombard and D.R. Duponteil, "PSK Transmission over a Nonlinear Satellite Channel," 1974 International Conference on Communication, Conference Record, Paper 36D, Minneapolis, MN, June 1974.

3. L. Lundquist et al., "Transmission of 4ϕ-PSK TDMA over Satellite Channels," IEEE Trans. Comm., Vol. 22, pp. 1354-1360, September 1974.

4. S. Yokoyama et al., "The Design of a PSK Modem for the TELESAT TDMA System," 1975 International Conference on Communication, Conference Record, Paper 44C, San Francisco, CA, June 1975.

5. W.H. Tranter, R.E. Ziemer, M. Fashano, "Performance Evaluation Using SYSTID Time Domain Simulation," 1975 International Conference on Communications, Conference Record, Paper 20A, San Francisco, CA, June 1975.

6. P. Balaban, "Statistical Evaluation of the Error Rate of the Fiber Guide Repeater Using Importance Sampling," Bell System Technical Journal, Vol, 55, pp. 7456-766, July 1976.

7. H.B. Poza, H.L. Berger and D.M. Bernstein, "A Wideband Data Link Computer Simulation Model," Comput. & Elect. Engineering, Vol. 5, pp. 135-149, Pergamon Press Ltd.,1978.

8. P.S. Angelo, M.C. Austin, M. Fashano, D.F. Horwood, "MSK and Offset Keyed QPSK through Band Limited Channels, "Fourth International Conference on Digital Satellite Communications, Montreal, Canada, October 1978.

9. W.C. Braun, "Analytical Simulation of Nonlinear Satellite Links," Proc. International Telemetering Conference, San Diego, CA, October 1980.

10. J.W. Modestino, G.R. Redinbo, K.R. Matis and A.L. Vickers, "Digital Simulation of Communication Links", Proceeding of the National Telecommunications Conf., Washington D.C., 1979.

11. W.R. Griese and P.K. Leong, "Interactive Communication systems Simulation Model - ICSSM", Proceedings of the international communications conference, Denver, CO, June 1981.

12. M.L. Steinberger, P. Balaban, and K. Sam Shanmugan, "On the Effect of Uplink Noise in Nonlinear Satellite Channels", Proceedings of the International Communications Conference, Denver, CO, June 1981.

13. K. Sam Shanmugan and P. Balaban, "A Modified Monte-carlo Simulation Technique for the Estimation of Error Rate in Digital Communication Systems", IEEE TRANSACTIONS ON COMMUNICATIONS, Vol 28, November 1980.

APPENDIX A. SIMULATION OF A SPREAD SPECTRUM MODEM OPERATING IN A JAMMING ENVIRONMENT

This appendix describes an approach to simulating the key features of a spread spectrum modem operating in a jamming environment. The proposed simulation will be carried out using two sets of models. One set of models will simulate generic type features that are used in a variety of spread spectrum modems. These features include PN spreading using Gold codes, MFSK modulator/demodulator, bandpass and low pass filtering, all pass channel with additive noise and interference, FH frame synchronizer, PN correlator/matched filter and delay-locked loop synchronizer. A second set of models or subroutines will be used to simulate the features of a particular modem such as the burst format, frame format, etc. The overall simulation will be done using ICSSM.

A block diagram of the features to be simulated is shown in figure A.1. Major features of the modem that will be simulated include the spread spectrum waveform generation, frame synchronization, bit level synchronization (tracking) and demodulation. The spread spectrum waveform will be simulated by using a random bit sequence generator for generating the input message sequence, a linear feedback shift register to generate Gold type PN spreading sequence, and a burst and frame formatter. Sampled values of the spread signal will be put on a MSK carrier and transmitted through a channel that is modeled as an all pass channel with additive Gussian noise and interference. The noisy values of the modulated signal will be stored for further processing. It must be emphasized here that lowpass equivalent of band pass signals/systems will be used in simulation so that sampling rate will be determined by the envelope bandwidth and not the carrier frequency. The sampling rate should be about 10 times the RF bandwidth.

Sampled values of the PN spread FH/MSK signal are processed to simulate the receiver. Assuming ideal dehopping at the receiver, the frame timing recovery operation is simulated using a bank of complex correlators and a combiner. While this simulation model may be different than typical frame synchronizer implementation for FH systems, it does provide an easy method for assessing the accruacy of the frame synchronizer as a function of the S/N and J/S ratio. A delay-locked tracking loop model is used to simulate the tracking aspect of the receiver. The PN despreader and the MSK demodulator model that will be used in the simulation are generic type models that can be used for other applications.

Completion of the software models for the functional blocks shown in Figure A.1. will provide DICEF with a set of library models for simulating a variety of spread spectrum systems. With a small additional set of functional models for a particular type of spread spectrum system (say the SEEKCOM Modem) simulations can be carried out via ICCSM to provide the following types of results.

(1) Probability of error as a function of J/N and S/N for various types of interference waveforms.

(2) Combined effects of different types of interference waveforms.

(3) Effects of interference and noise on frame and bit synchronizing.

This proposed development will serve the following needs:

(1) To establish a frame work for modeling major features of spread spectrum modems, and
(2) To help assess the susceptibility of a specific spread-spectrum modem operating in a jamming environment.

Cross validation of the simulation results with hardware measurements will considerably enhance the credibility of using simulations, particularly among "hard-nosed" analysists who are skeptical about simulations.

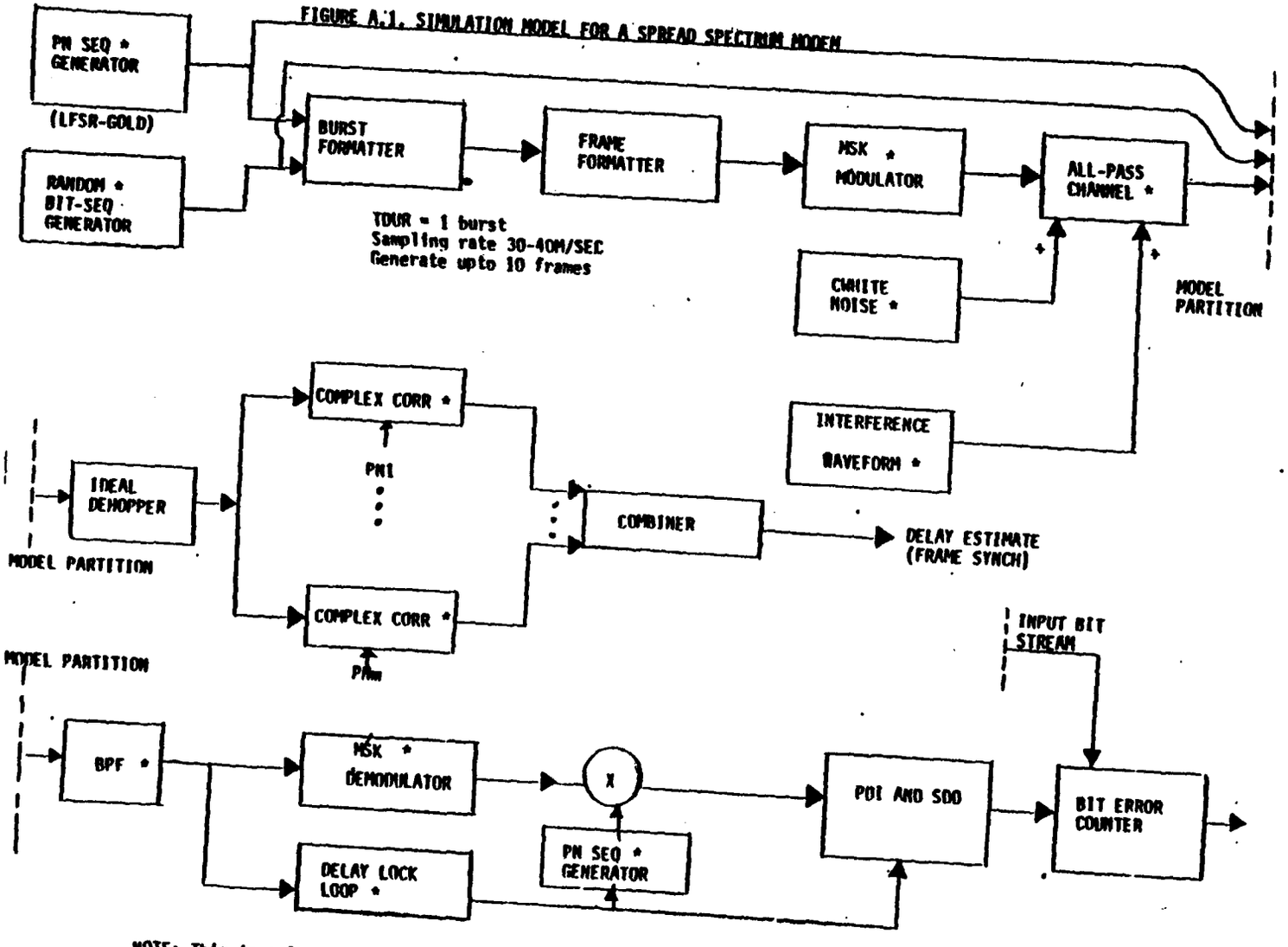


FIGURE A.1. SIMULATION MODEL FOR A SPREAD SPECTRUM MODEM

NOTE: This is a lowpass equivalent simulation; * denotes general models

1982 USAF-SCEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

AFTERBURNING SUPPRESSION KINETICS IN ROCKET EXHAUST

Prepared by: Trilochan Singh

Academic Rank: Associate Professor

Department and University: Mechanical Engineering
Wayne State University
Detriot MI 48202

Research Location: Air Force Rocket Propulsion Laboratory
Propulsion Analysis Division
Plume Technology Branch
Edwards AFB CA 93523

USAF Research College: Dr Jay Eversole

Date: August 13, 1982

Contract No: F49620-82-C-0035

AFTERBURNING SUPPRESSION KINETICS IN ROCKET EXHAUST

by

Trilochan Singh

ABSTRACT

Two computer codes have been used to study the suppression of afterburning phenomena during the summer of 1982. PLFLAME code (which sovles a set of equations governing a premixed laminar flow) was used to simulate the temperature and species profiles for a fuel rich H_2-Air flame. The results of PLFLAME code show that the addition of candidate suppressants (K, HBr) lower the peak flame temperatures. The steady state atomic hydrogen mole fraction decreases as the percentage of K increases. The trend is reversed with the increase in HBr mole fraction.

Boat code was used to simulate the mixing and chemical processes that taken place as the exhaust jet mixes with outside stream. The results of BOAT code show that the addition of K to the unburnt stream delays the onset of ignition at some radial locations depending upon the chemical composition of the jet.

Additional runs are required to identify the kinetic mechanisms responsible for the phenomenon and design an experiment to verify the results of the simulation.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the assistance of the Air Force Systems Command, Air Force Office of Scientific Research for the sponsorship and financial support for this project and the management of the USAF Rocket Propulsion Laboratory for providing the computer facilities. The author is thankful to Mr Wilbur Andrepont, Dr Jay Eversole, Dr Dave Weaver, Dr Pat Saatzer and Dr Bob Rhodes for helpful discussions at the various stages of this study. The author is also thankful to Susan Hulsizer, Julie Casper, for their help in preparing the manuscript. Finally, the author would like to express appreciation to the TOD Branch and ITT for their assistance with computer operations.

AFTERBURNING SUPPRESSION KINETICS IN ROCKET EXHAUSTS

I. INTRODUCTION

The exhaust gases from the nozzle of a rocket motor usually contain significant proportions of unburned fuel. This fuel mixes turbulently with ambient air as the exhaust jet expands and may burn causing a substantial elevation of temperature in the exhaust plume. The secondary combustion is affected by factors such as velocity and altitude of the missile, motor thrust level, the pressure, temperature at the nozzle exit plane. The addition of certain species such, as K, K_2SO_4, HBr, etc., have been observed to inhibit the afterburning process. The inhibition of afterburing process is of great interest to the US Air Force. However, the kinetic mechanism for the afterburning suppression is not fully understood. A research program has been initiated at the Rocket Propulsion Laboratory to investigate the afterburning suppression kinetics by Dr Jay Eversole. The fundamental research objective will be to identify intermediate species that are critical to the afterburning suppression phenomena.

II. APPROACH AND OBJECTIVE

One possible experimental approach being considered is to run a preburner in a fuel rich mode and allow the exhaust products to diffusively mix with a surrounding stream of air as shown in the Figure I. Candidate species as suppressants would be introduced in the preburner. The species concentrations, temeprature and local velocity will be measured as a function of radial and axial position.

It is essential that the experimental approach be supported by a computational model. The objective of my summer assignment is to study the feasibility of this

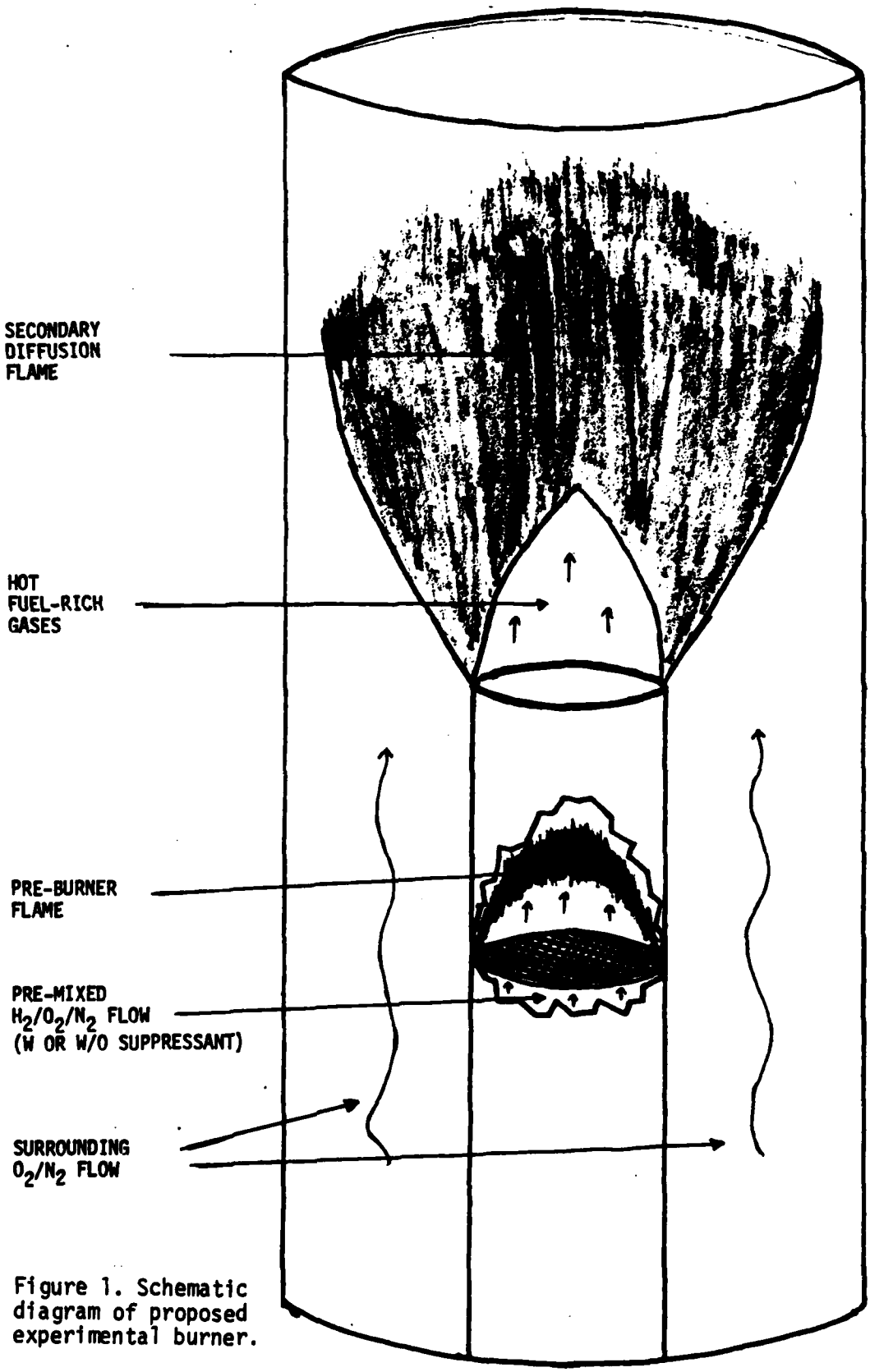


Figure 1. Schematic diagram of proposed experimental burner.

experiment by using an appropriate set of computer programs. A review of the literature[1-5] shows that two computer codes, BOAT and PLFLAME, (developed by the JANNAF Community) with some modifications may be used to simulate the proposed experiment. The main features of these two computer codes are summarized below.

III. DESCRIPTION OF PLFLAME PROGRAM

PLFLAME code was obtained from Dr Terence P. Coffee of the US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland in mid-June. PLFLAME code simulates the combustion process occuring in the laminar, premixed, one dimensional steady state flame. PLFLAME code is divided into four sequential parts.

The first part requires thermochemical data (such as given in the JANNAF thermochemical tables) and the transport properties data (such Leonard-Jones potential parameters) for all the species in the reaction scheme as input. This first part of the PLFLAME code produces a data file containing a set of coeffients for the various polynomials fits used for calculating the transport properties of the species as a function of temperature. This data file is used as input to the second part of the PLFLAME code. The mole fraction for the incoming stream, the inlet temperature and the pressure are given as input by the user for the PLFLAME code, part II. PLFLAME Code I also produces three subroutines RT, RATE and F which are used by part II and part III. These subroutines calculate the rates of the forward reactions and the backward reactions involved in the proposed kinetic mechanism.

PLFLAME code II generates a set of data containing the inlet boundary conditions and the initial boundary conditions and a first-order guess for the composition. This is used as a starting point by the PLFLAME code III, which numerically solves a set of

coupled differential equations governing a laminar, premixed, one dimensional, steady state flame. The basic procedure is to integrate the equations in time until the steady state solution is reached. Reference 2 describes in detail the partial differential equations governing the flames and the method of solution used in the code. PLFLAME code produces the output for the temperature distribution and the mole fraction distribution of the various species.

IV. DESCRIPTION OF BOAT PROGRAM

BOAT code has been developed by S. M. Dash and Harold S. Pergament for NASA and the Army[3-5]. BOAT code accounts for the detailed turbulence and thermochemical processes occuring in the mixing layer formed between a jet exhaust and its surrounding external flowfield regions in an overlaid interactive manner. BOAT code combines the best features of the following three codes to solve this flowfield.

(1) The generalized treatment of thermochemical processes and implicit/explicit streamline/integration procedure used in the Low Altitude Plume Program (LAPP)[6] has been employed in BOAT code.

(2) The shear layer distribution, grid distribution and growth rules used in the GENMIX[7] code have been incorporated in the BOAT code.

(3) The overlaid procedure for describing the variable edge conditions used in the GASL patched system for the detailed analysis of inviscid/shock and mixing/afterburing process in rocket exhaust plumes[8] has been adapted in BOAT code.

Reference 9 summarizes the main features of the BOAT code and the governing equations. The description of the two codes show that PLFLAME code would be suitable for simulating the flame generated by the gases flowing through the internal tube (Fig. 1). The BOAT code would be suitable for simulating the processes occuring in the mixing layer as the exhaust products from the preburner mix diffusively with the surrounding gases.

V. DIFFICULTIES IN ACTIVATING THE COMPUTER PROGRAMS

A number of difficulties were encountered in activating the PLFLAME code on the AFRPL computer system. Some of these were related to in the job control cards, but most of the difficulties were caused by errors in the two subroutines (RT, RATE) generated by part I of the PLFLAME code and subsequently used by part II and part III of the PLFLAME code. Two subroutines (DEC and SOL) which are part of the computer library at Aberdeen are not available in the AFRPL computer library. It took over four weeks to eliminate all the errors and add the two subroutines. The program was validated by comparing the output for a sample problem for ($H_2/O_2/N_2$) generated by the AFRPL computer system with output generated at the Aberdeen computer system.

BOAT code was activated during the 2nd week after my arrival. The code was validated by comparing the output for a sample problem generated by the AFRPL computer system with the output given in the user's manual[4]. A number of runs were made for the BOAT code to check the sensitivity of the program to the changes in the input parameters. The following iput parameters were changed: input chemical composition, velocity of the exhaust jet, (5 fps to 500 fps), Ambient Stream Mach number, (0.01 to 0.5), nozzle radius, (0.15 ft to 2.54 ft), jet temperature, (300-1200K), and print stations. The program did not work when the change in the print internval was not specified (assuming equal interval throughout). The BOAT output was not satisfactory when the jet velocity and the ambient stream velocities were too close to each other. At other combinations the program worked satisfactorily.

VII. RESULTS OF SUPPRESSION MECHANISMS USING PLFLAME

After establishing satisfactory sample problem outputs for the two programs it was decided to study the effect of adding two candidate suppressants: K (potossium)

and HBr to H_2-O_2-N_2 flames as suggested by the studies of Jensen[10-12] and Dixon-Lewis[13]. Using these two suggested mechanisms, PLFLAME code was run for different inlet gas compositions (mole fractions) given in Table I.

TABLE I

	H_2	O_2	N_2	HBr	K	
1.	.599	.089	.312	-	-	(60%H_2-60% Air)
2.	.576	.086	.312	.04	-	(4% HBr)
3.	.552	.0773	.2407	.08	-	(8% HBr)
4.	.576	.0806	.3054	-	.04	(4% K)
5.	.552	.0773	.7907	-	.08	(8%)

The system pressure was assumed to be atmospheric. the flow velocity was kept at 100 cm/second and the burner surface temperature was assumed to be 600K for these computer runs. Table 2 lists the reactions compsosing the suppression mechanisms postulated for K and HBr addition and provides the forward and reverse reaction rate coefficients used in both PLFLAME and BOAT. The species mole fraction and the temperature distribution as a function of distance above the burner surface are shown in the Fig. 2, 3, and 4. The calculated peak flame temperature, steady state atomic hydrogen mole fraction for the various runs are summarized in Table 3. A review of the figures 2-4 and Table 3 show that the peak flame temperature and the flame length (distance over which significant changes take place) decreases as the percentage of HBr or K increases. The mole fraction of atomic H decreases as the concentration of K increases. This trend is reversed as the mole fraction of HBr increases in the unburnt gas mixture.

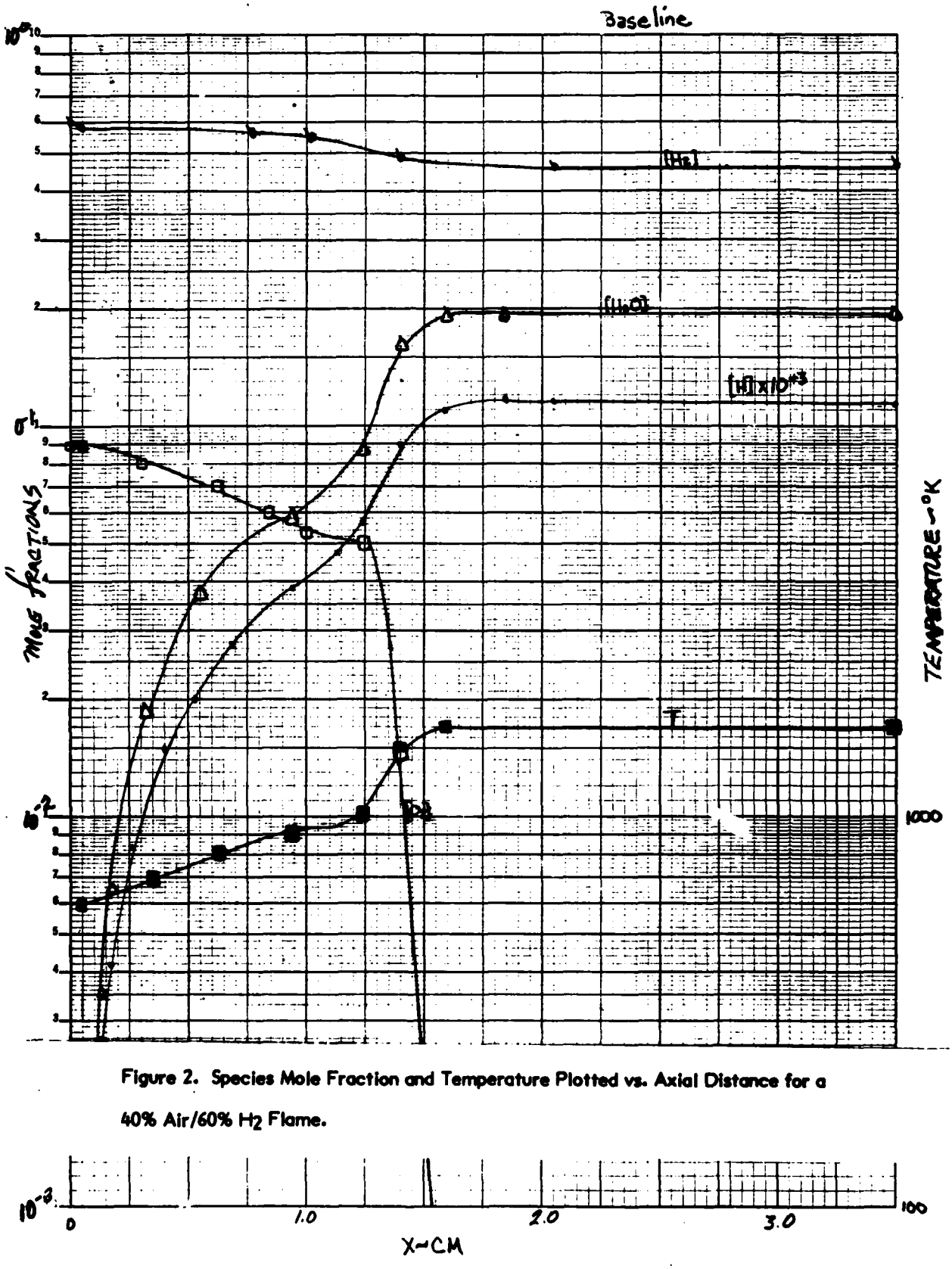


Figure 2. Species Mole Fraction and Temperature Plotted vs. Axial Distance for a 40% Air/60% H_2 Flame.

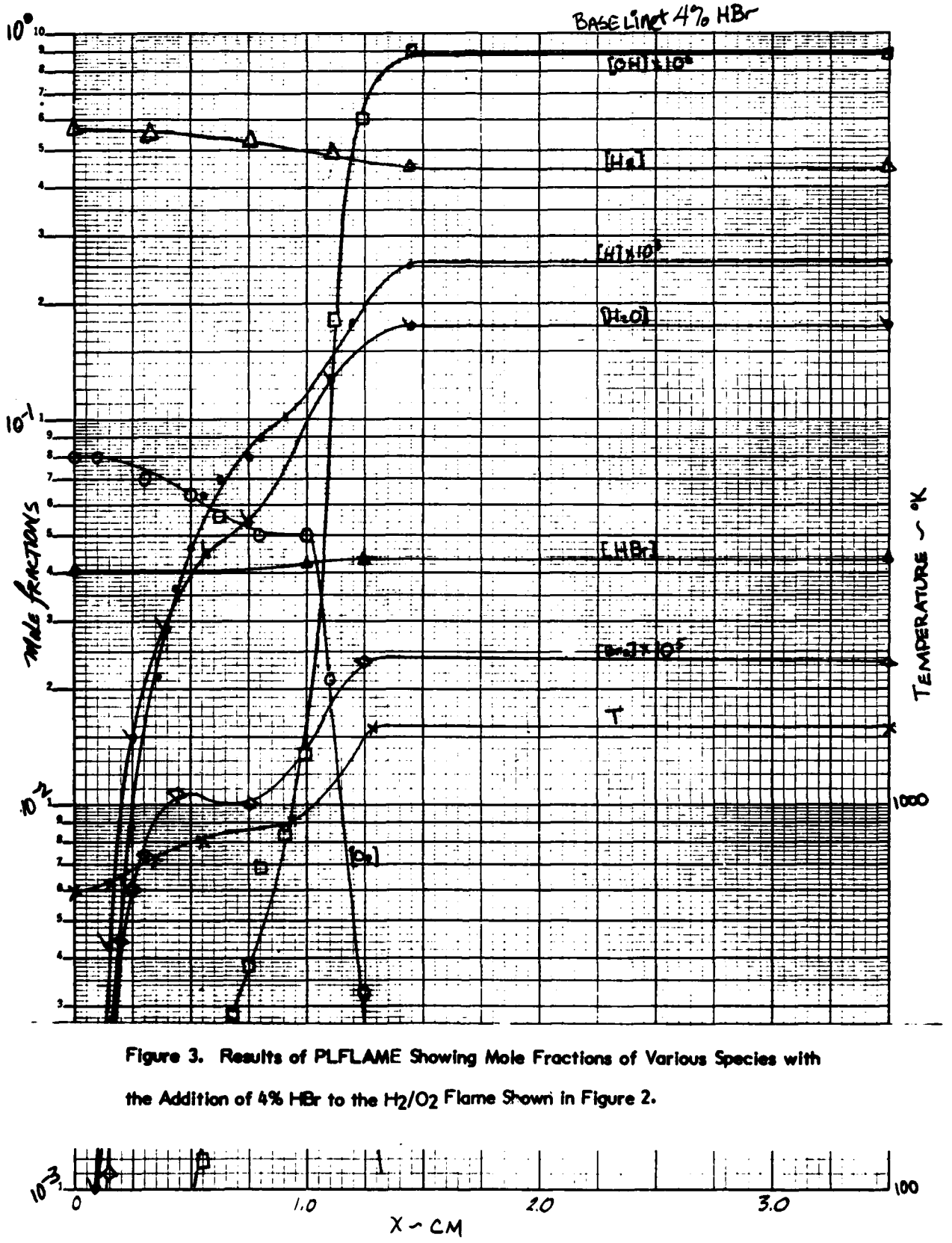


Figure 3. Results of PLFLAME Showing Mole Fractions of Various Species with the Addition of 4% HBr to the H_2/O_2 Flame Shown in Figure 2.

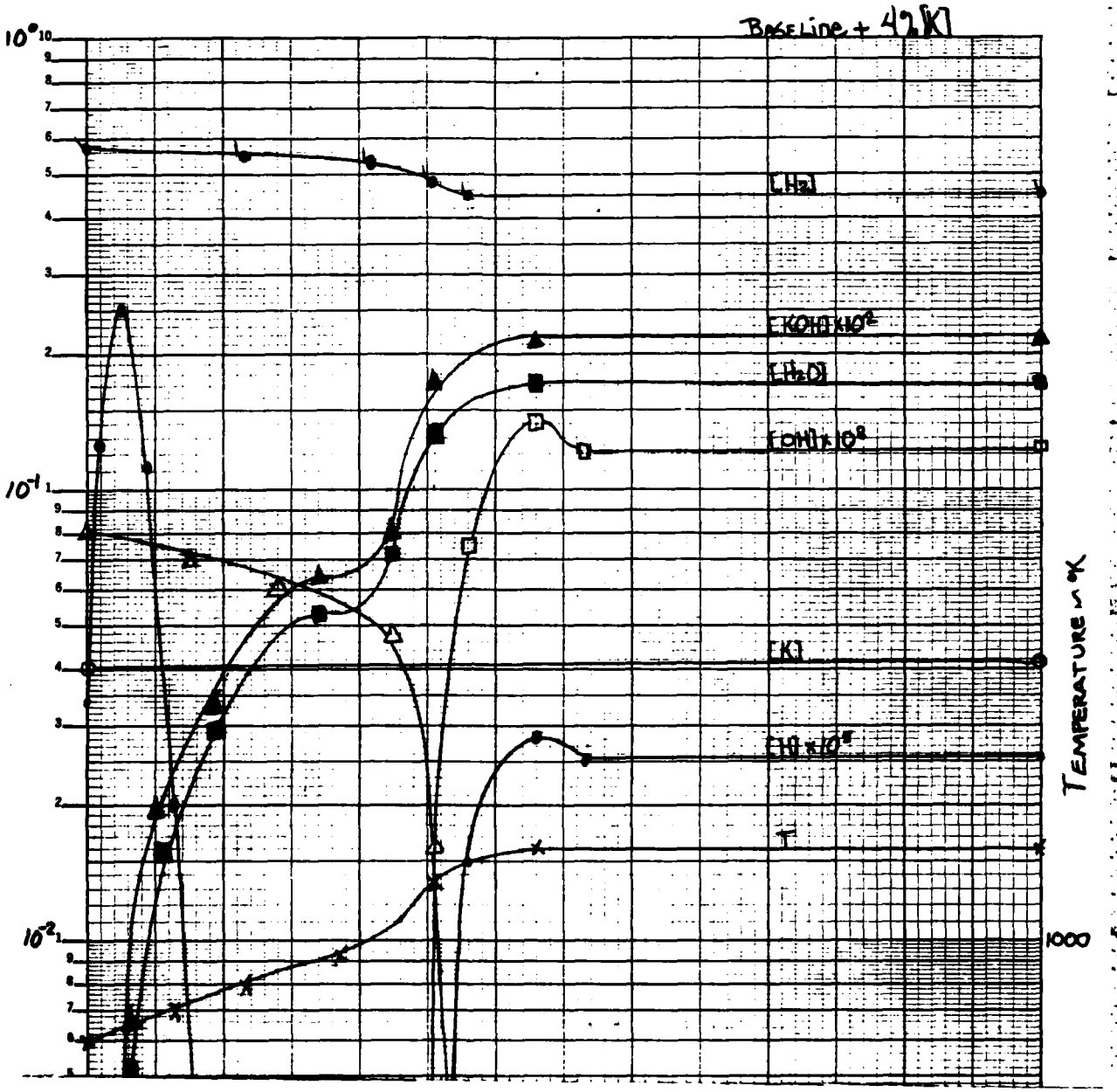


Figure 4. Results of PLFLAME Showing Mole Fractions of Various Species with the Addition of 4% K to the H_2/O_2 Flame Shown in Figure 2.

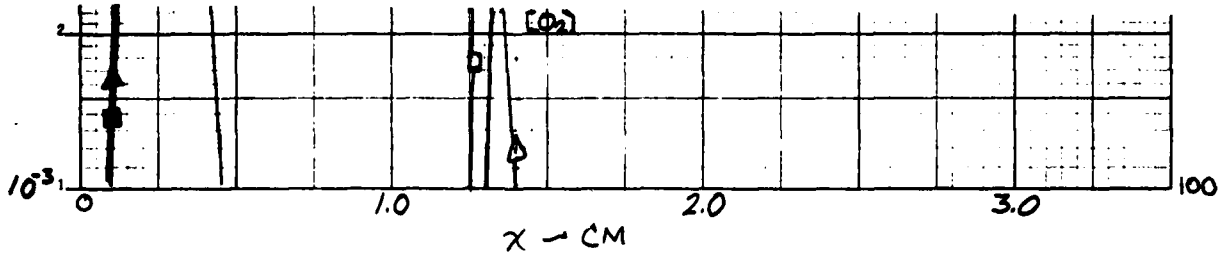


TABLE 2

INHIBITION REACTIONS

	$k=AT^n\exp(E/T)$	A	n	E
18	K + H20 = KOH + H	6.02E+13	0.	-2.04E+04
	KOH + H = K + H20	9.54E+17	-1.52E+00	-4.12E+03
19	K + OH + M = KOH + M	3.61E-04	1.00E-01	0.
	KOH + M = K + OH + M	2.39E+01	-1.25E+00	-4.36E+04
20	H + HBR = H2 + BR	5.30E+13	0.	-1.50E+03
	H2 + BR = H + HBR	6.43E+13	1.03E-01	-9.83E+03
21	BR + HBR = H + BR2	2.70E+14	0.	-2.20E+04
	H + BR2 = BR + HBR	3.03E+16	-5.06E-01	-1.17E+03
22	BR + BR + H2 = BR2 + H2	1.30E+17	-7.00E-01	0.
	BR2 + H2 = BR + BR + H2	4.99E+18	-1.03E+00	-2.29E+04
23	H + BR + H2 = HBR + H2	9.77E+17	-7.00E-01	0.
	HBR + H2 = H + BR + H2	3.41E+17	-5.34E-01	-4.37E+04
24	BR + HO2 = HBR + 02	3.40E+12	0.	0.
	HBR + 02 = BR + HO2	6.05E+11	2.03E-01	-2.02E+04

TABLE 3

CALCULATED PEAK TEMPERATURE

FLAME LENGTH

(H) MOLE FRACTION

RUN NO	UNBURNT GAS COMPOSITION	TEMP	FLAME LENGTH	STEADY STATE (H)-MOLE FRACTION
1.	60% H_2-40% Air	1716	3.65 CM	1.1667E-04
2.	4% HBr	1588	2.837	2.539E-04
3.	8% HBr	1537	2.340	3.1808E-04
4.	4% K	1611	3.23	2.532E-07
5.	8% K	1580	2.96	1.5927E-07

A comparison of these PLFLAME code results with those of Dixon-Lewis[13] shows the following differences:

The steady state H atom mole fraction decreases as the % HBr increases for the Dixon-Lewis[13] study where as the results of the PLFLAME code show the reverse trend.

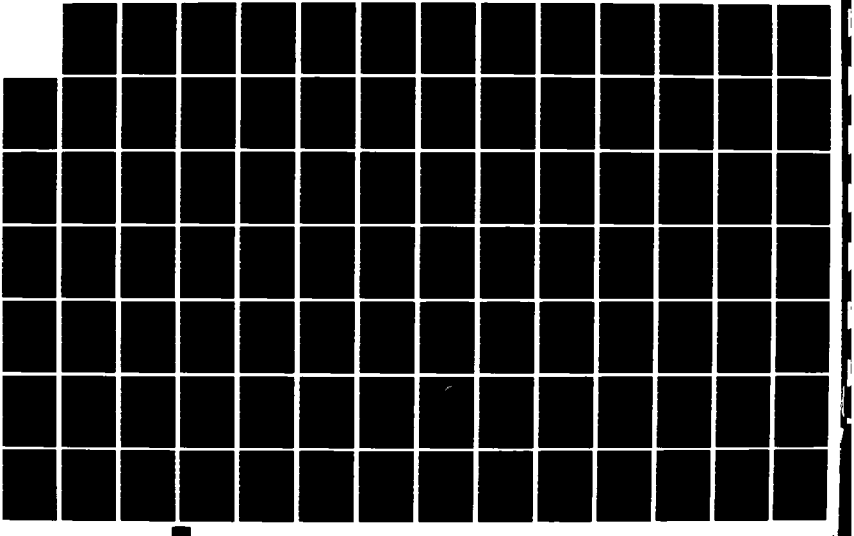
The peak temperature does not change in the Dixon-Lewis[13] study.

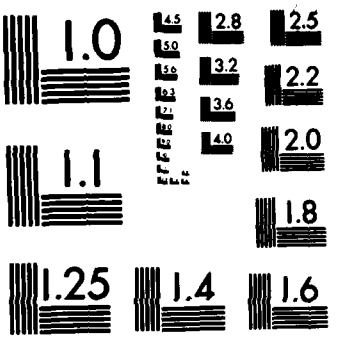
The flame increases from 3 mm to 4 mm as 4% HBr is added to 60% H_2 Air flame.

The reasons for the differences in results have not been investigated, but these can be caused by factors such as the differences in the assumptions (steady state assumption, partial equilibrium assumption, etc.), in the formulation of the equations, transport properties differences, etc., Dixon-Lewis[13] have not compared the results of calculation with experimental data. All these factors need additional study.

A review of Table 3 shows that the peak temperature decreases from 1711 K to 1580 K as the concentration of K increases from 0 to 8%. The concentration of atomic (H) decreases as K increases as shown by Figures 2 and 4. This observation is in agreement with the experimental data of Jensen[11] which shows a decrease in the atomic (H) concentration. The calculated length of flame (about 3cm) is close to that of Jensen[11]. The experimental data of Jensen could not be simulated during the summer 1982 for lack of information (such as burner suface temperature, flow velocity, transport property, etc). The PLFLAME code should be run for the experimental conditions used by Jensen and the reasons for the differences (if any) should be investigated.

AD-A130 770 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM (1982) 8/11
RESEARCH REPORTS VOLUME..(U) SOUTHEASTERN CENTER FOR
ELECTRICAL ENGINEERING EDUCATION INC S..
UNCLASSIFIED W D PEELE ET AL. OCT 82 AFOSR-TR-83-0614 F/G 5/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

The postulated reaction mechanism for the inhibition of the $H_2/O_2/N_2$ flame for K or HBr addition is shown in Table 2. The net result of both mechanism is the consumption of H atoms by KOH or HBr species. The decrease in the availability of H atoms to carry out the chain reaction of H_2-O_2 system is the primary reason for H_2/O_2 flame inhibition. This trend for decrease of H-atom is not shown for HBr addition by the PLFLAME output. The reasons for this difference need to be investigated

From the preliminary resluts of the effect of adding (K) and HBr to H_2-air flames, it appears that the PLFLAME code is working satisfactorly. The code needs to be calibrated by comparing the calculated results with the experimental data of various workers and then use it to plan the conditions for the proposed experiment.

VII. RESULTS OF SUPPRESSION MECHANISMS USING BOAT

BOAT code was run for 2 sets of conditions shown in Table 4.

TABLE 4

BOAT CODE WAS RUN FOR THE FOLLOWING SET OF CONDITIONS

	RUN	COMPOSITION	INPUT
	1	60% H_2-40% Air	PLFLAMES Output
SET I	2	4% K	"
	3	8% K	"
	4	60% H_2-40% Air	Unburnt Composition
SET 2	5	4% K	"
	6	8% K	"

The calculated temperature profile at two radial stations and various axial positions is shown in Figures 5 and 6 for set 2. A review of the results show that for set I with r = 5 (which is 5th radial starting out of II total) ignition is delayed for 4% K and 8% K flame. At r = 7 the trend is reversed. There is no ignition for 0% K flame but there is rapid ignition for 4% K and 8% K flame. The reason for this behavior is not understood.

A second set of computer runs with the BOAT code was made (see Table 5). It was assumed that the jet composition is the unburned composition (60% H_2 - 40% air with no K, 4% K, and 8% K). The temperature profiles at two radial positions (r=5, 7) and various axial positions is shown in Figures 7 and 8. These plots show that ignition is delayed for 4%K addition and there is no ignition for 8% K addition for the first 2 ft (about 12 nozzle radii). The reaction rates, reaction mechanism and kinetic turbulence parameters were unchanged. The reasons for this behavior needs further investigation which could not be done during the limited summer time.

VIII. SUMMARY AND CONCLUSIONS

The limited work done during the summer leads to the following points:

- PLFLAME code and BOAT code are working satisfactorily.
- The addition of HBr and K lower the calculated peak flame temperature as shown by PLFLAMES output.
- The concentration of (H) atoms plays a significant role in the inhibition of flame according to the postulated mechanisms.
- The calculated concentration of H atoms decreases as the % of K increases this trend is reversed for the addition of HBr.
- The results of BOAT code show that ignition is delayed with increasing % of

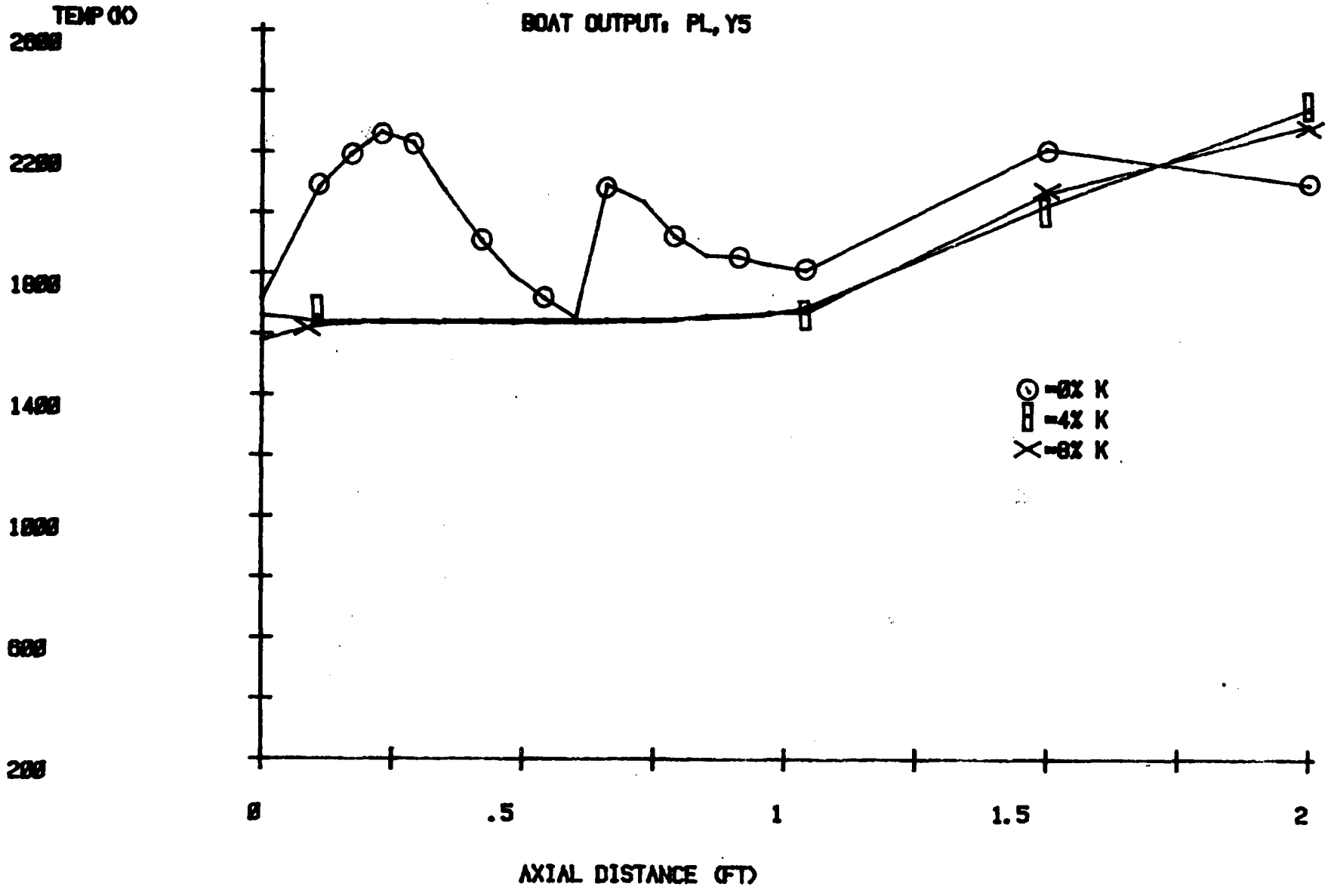


Figure 5. Calculated temperatures from BOAT program using the output of PLFLAME with 0.0, 5.0, and 8.0 percent K added (See figures 2 and 4) used as initial conditions. Temperatures are plotted versus axial station for the fifth radial station.

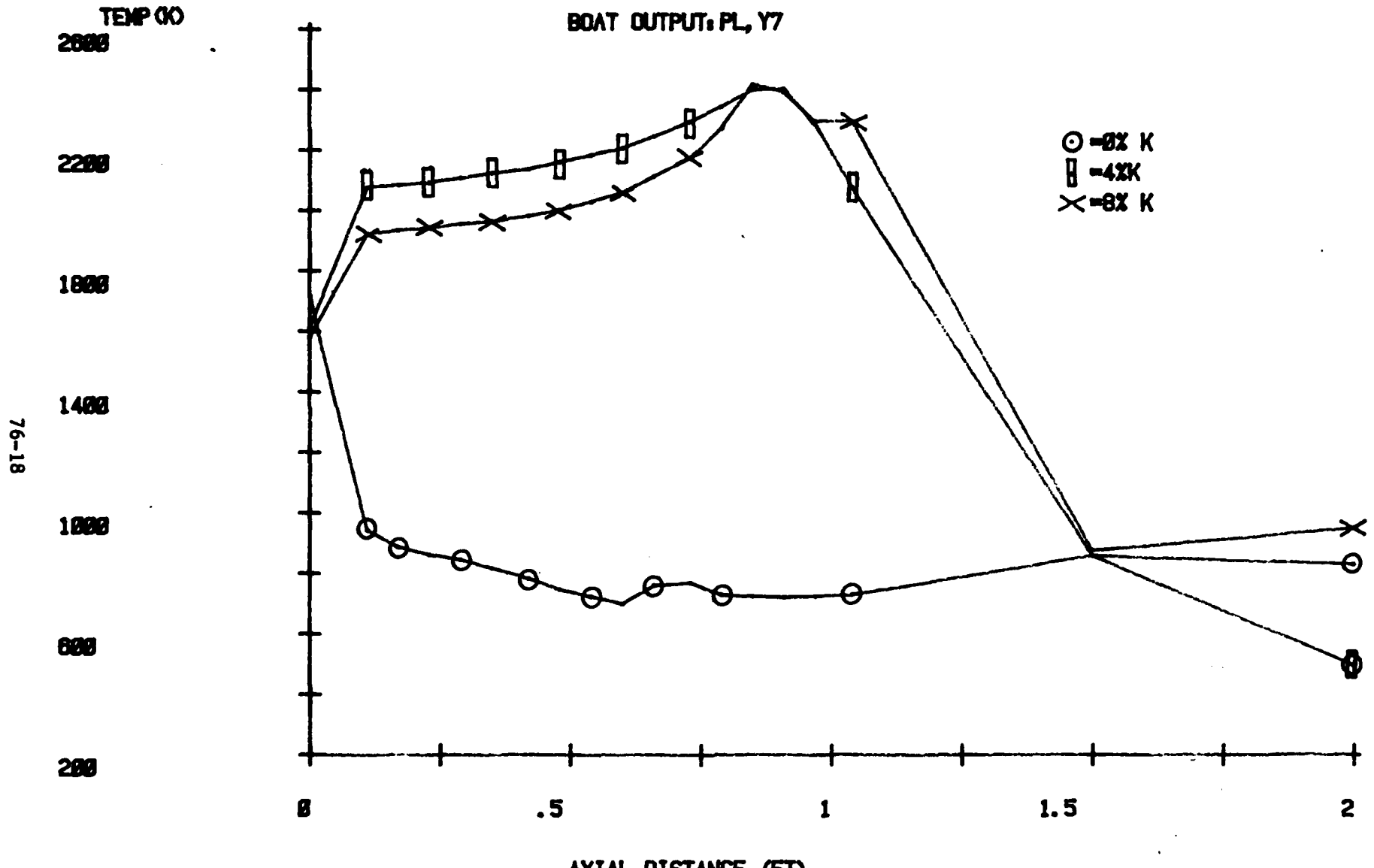


Figure 6. Calculated temperatures from BOAT program under the same conditions as Figure 5, except the temperatures are from the seventh radial station

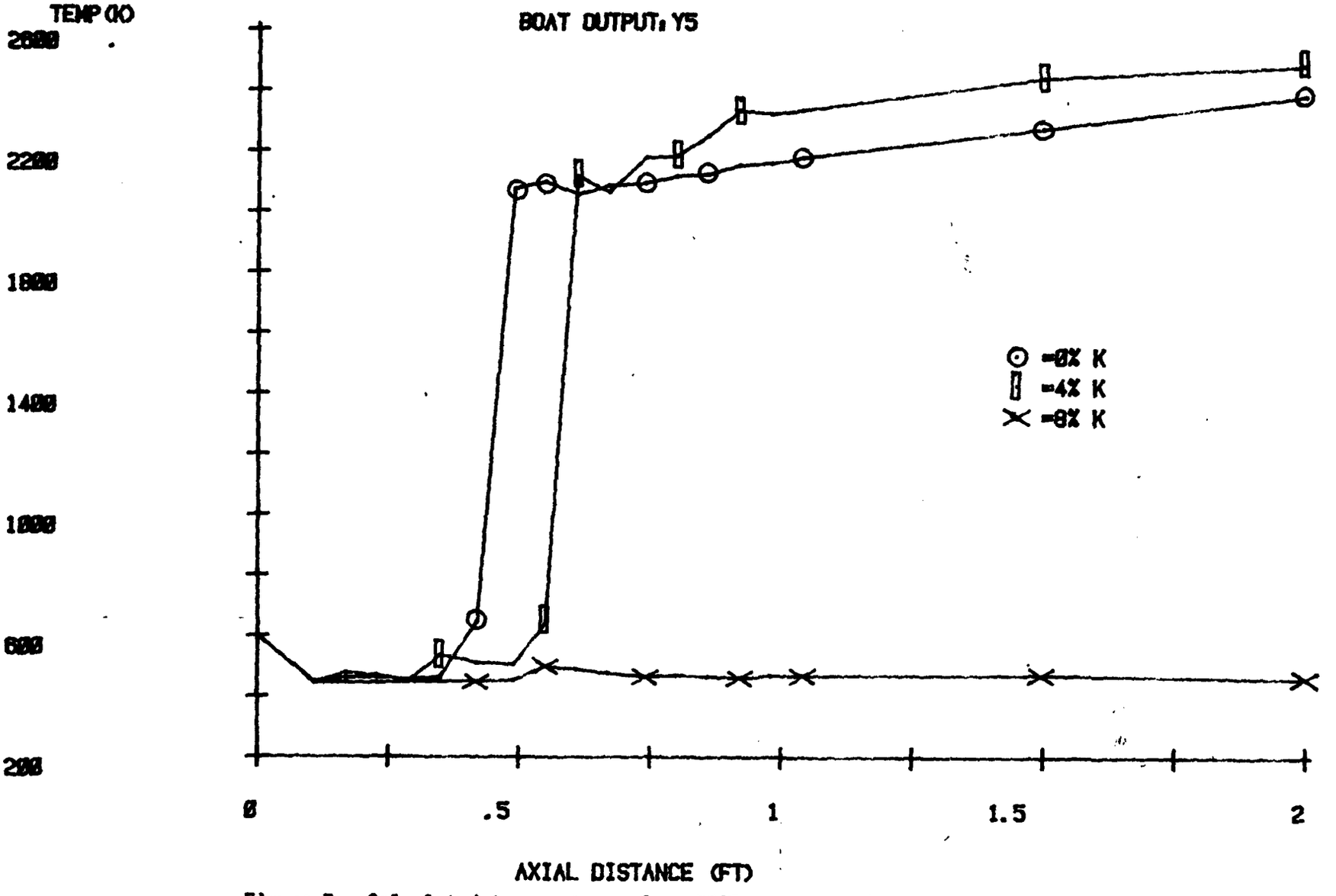


Figure 7. Calculated temperatures from BOAT program using an unburned mixture of gases including K as initial concentrations at a temperature of 600K. Temperatures of the fifth radial station are plotted versus axial distance.

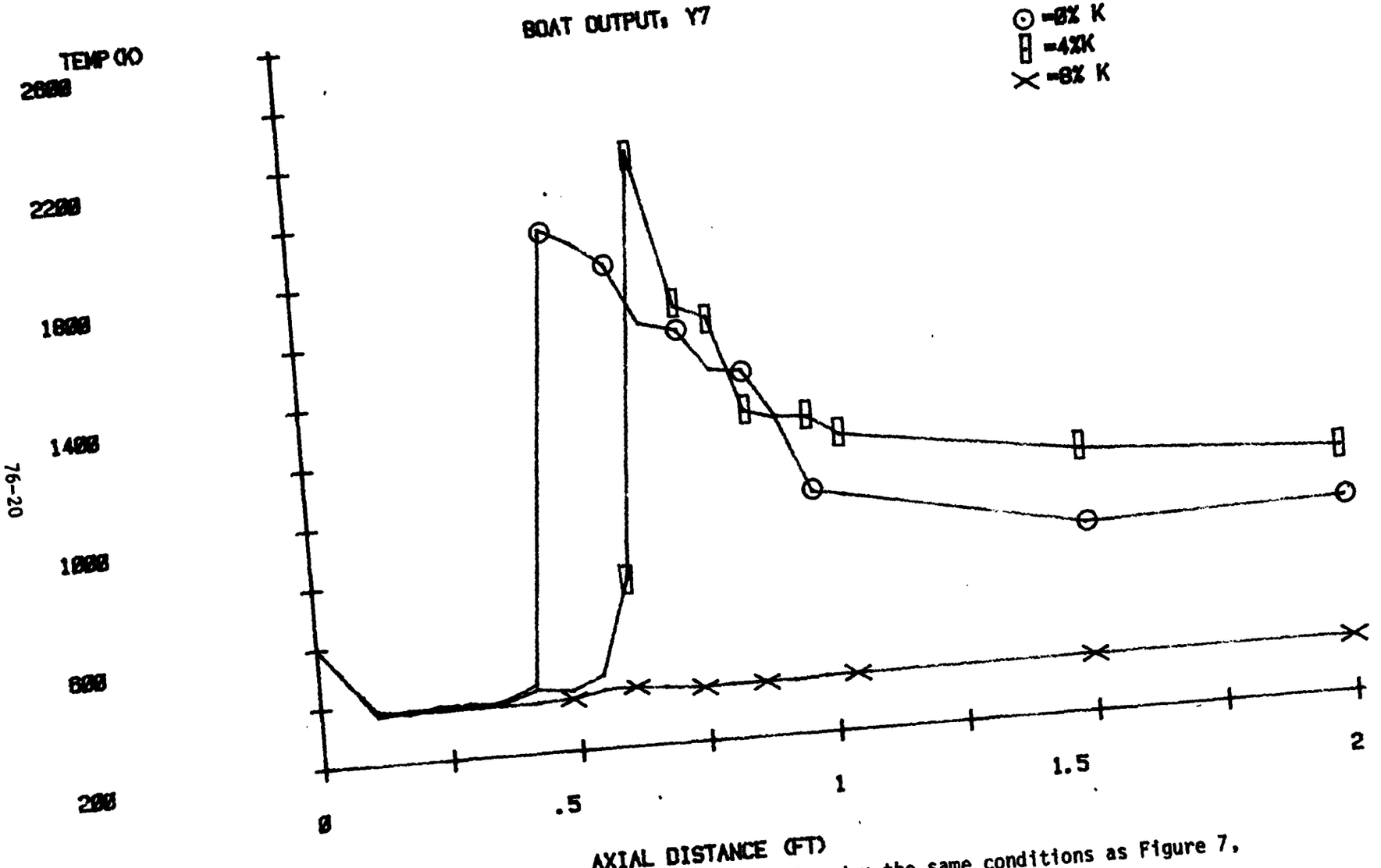


Figure 8. Calculated temperatures from BOAT under the same conditions as Figure 7, except the temperatures are from the seventh radial station.

K when the input is the unburnt composition. This trend is reversed when the output from PLFLAMES is used as input to the BOAT. The reasons for this behavior is not understood at this time.

IX. RECOMMENDATIONS

- PLFLAME code and BOAT code should be calibrated against the available experimental data in the literature before using it for the design of the proposed experiment.
- The output of PLFLAME code should be checked against a set of equilibrium calculations for the exit plane of PLFLAME3 code.
- After calibration and equilibrium calculations PLFLAME code should be used to predict the effect of adding suppressant species such as K and HBr on the flame structure then results would enable one to specify the conditions such as F/A, (pressure, temperature, etc.) for the unburnt gases going into the preburner.
- BOAT code should be run for different combinations of

  Fuel air ratio

  initial velocity, temperature

  nozzle size

  outer stream conditions

The output of PLFLAME code should be used as input to simulate the condition in the mixing layer between the inner jet and the outside stream. The results would predict the condition in the mixing layer between the inner jet coming out of preburner and the outer concentric stream for the proposed experiment.

X. REFERENCES

1. Heimerl, J. M., and Coffee, T. P., "The Detailed Modeling of Premixed Laminar Steady-State Flames," Combustion and Flame 39: 301-315C1980).

2. Coffee, T. P., "A Computer Code for the Solution of the Equations Governing a Laminar, Premixed One-Dimensional Flame," Report ARBRL-MR-03165, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD 21005, 1982.

3. Dash, S. M., and Pergament, H. S., "JANNAF Standard Plume Flowfield Model (SPF)," U.S. Army Missile Command, Technical Report No. RD-CR-82-9, April 8.

4. Dash, S. M., Pergament, H.S., Thorpe, R. D., Abuschowski, S. A., and Hussain, J. C., "Operational Instructoins for a Prelimnary Version of the JANNAF Standard Plume Flowfield Model (SPF/I)," U.S. Army Missile R&D Command, Contract No. DAAK40-78-C-0124, ARAP Report Report #415, June 1980.

5. Dash, S. M., and Pergament, H. S. "A Computational Model for the Prediction of Jet Environment in the Vicinity of Nozzle BOAT TAILS (BOAT Code)," NASA Contract No. NASI-14794, Report 3075, Dec 1978.

6. Mikatarian, R. R., Kau, C. J., and Pergament, H. S., "A Fast Computer Program for Nonequilibrium Rocket Plume Predictions," AFRPL-TR-72-94.

7. Patankar, S. W., and Spalding, D. B., "Heat and Mass Transfer in Boundry Layers," Inter Text Books (London), 1970.

8. Dash, S., Boccio, J., and Wheilerstein, G., "A Computational System for the Prediction of Low Altitude rocket Plume Flowfields, Vol I-III," GASL TR-239, General Applied Science Lab, Inc., Westbury, NY, Dec 1976.

9. Dash, S. M., and Pergament, H. S., "A Computational System for the analysis ofMixing/Chemical/Shock Processes in Supersonic Internal and Exhaust Plume Flowfields," AIAA-80-1255.

10. Jensen, D. E. and Jones, G. A., "Kinetics of Flame Inhibition by Sodium," J. Chem Soc. Faraday Trans. I, 1982, 78.

11. Jensen, D. E., Jones, G. A., and Mace, C. H., "Flame Inhibition by Potassium," J. Chem. Soc. Faraday I, 1979, 75, 2377-2385.

12. Jensen, D. E., "Alkali-Metal Compounds in Oxygen Rich Flames," J. Chem. Soc. Faraday, Trans I, 1982, 78.

13. Dixon-Lewis, G., "Mechanism of Inhibition of Hydrogen-Air Flames by Hydrogen Bromide and its Relevance to the General Problem of Flame Inhibition," Combustion and Flame 36: 1-14 (1979).

1982 USAF-SCEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

ON ESTIMATING PROBABILITY OF KILL IN WEAPON EVALUATION

Prepared by:	Dr. B. D. Sivazlian
Academic Rank:	Professor
Department and University:	Department of Industrial and Systems Engineering The University of Florida
Research Location:	Armament Laboratory, Analysis Division, Weapon Evaluation Branch (DLYW), Eglin AFB FL
USAF Research Colleague:	Mr. Charles A. Reynolds
Date:	August 20, 1982
Contract No:	F49620-82-C-0035

ON ESTIMATING PROBABILITY OF KILL

IN WEAPON EVALUATION

by

B. D. Sivazlian

ABSTRACT

The problem of estimating the mean and standard deviation of the probability of kill when evaluating weapons in the presence of uncertainty in the input parameters is investigated . Kill produced by either weapon blast or weapon fragmentation is studied. Three different methods for carrying on the estimation procedure are analyzed. Method 1 relies on the assumption that the input parameters are uniformly distributed over given ranges. Method 2 assumes that the probability of kill function can be expanded as a Taylor's series in terms of the input parameter. A third method, closely related to Method 2, is also presented. The two situations when aiming error is absent and when aiming error is present are considered. Closed form expressions are provided whenever possible and suggestions for further research are offered.

Acknowledgement

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to spend an interesting and challenging summer at the Armament Laboratory, Eglin AFB, Florida. He would like to acknowledge the laboratory, in particular the Weapon Evaluation Branch, for its hospitality and excellent working conditions.

Special thanks are due to Mr. Charles A. Reynolds for suggesting this area of research and for his collaboration and guidance. Helpful discussions with Mr. James L. Thoreen, Mr. Marvis R. Adams, Mr. Jerry P. Bass, Mr. Allen L. Blackford, Mr. Patrick C. Coffield, Mrs. Katherine H. Douglas, Mr. Joel B. Knight, and Mr. Daniel A. McInnis are acknowledged.

I. INTRODUCTION:

In evaluating the effectiveness of a weapon, too much emphasis and reliance are put on single effectiveness numbers such as probability of kill P_k. For example, in comparing two weapon systems, the one with the highest P_k has traditionally been the system with the highest effectiveness. Little attention has been given to the variability associated with the value of P_k.

Basically, the value of P_k for a target point is determined from two sources:

(1) Vulnerability and warhead effectiveness measures. This usually provides an estimate of the capability of a weapon to destroy a target. Sometimes human judgement enters in these measures. For example, for fragment sensitive targets, the Carleton damage function may be used to measure the value of P_k at an arbitrary point (x, y) in the plane given that the weapon bursts at the origin (0,0). This damage function has the form

$$\exp\left[-\left(\frac{x^2}{R_x^2}+\frac{y^2}{R_y^2}\right)\right]$$

where R_x and R_y are parameters known as weapon radii which are determined experimentally.

(2) Delivery accuracy measures to the target. For example, a combination of aiming error and ballistic error contribute towards a net measure of delivery accuracy. Usually, these errors are described by Guassian distributions which may or may not incorporate the effect of bias. For example, under the assumption that the aiming errors are unbiased and under the absence of ballistic error, the standard deviations σ_x and σ_y of the aiming errors along the range axis (x - axis) and deflection axis (y - axis) provide a measure of the delivery accuracy.

It is customary to combine the information from these two sources to arrive at a measure of P_k. For example, for fragment sensitive targets assuming unbiased aiming error, the value of P_k would depend on the four input parameters R_x, R_y, σ_x and σ_y. Thus, a measure of the variance of P_k should be dictated by the statistical characteristics of each of the four input parameters.

Very little work appears to have been done in this area (see e.g. [2], [5]). The typical weapon effectiveness studies such as [1], [3], [4], [6], [7], do not discuss this particular aspect of the problem and provide only methodologies to arrive at a single estimate value of P_k.

The motivation for developing a methodology to compute estimates of the mean and variance of the probability of kill P_k is threefold. First, it allows the setting of confidence intervals on the P_k value. Even if such confidence intervals are not accurate, it still provides adequate information on how P_k behaves statistically in the presence of estimation error of the various input parameters. Second, it gives the decision maker a necessary information to arrive at a rationale for comparing and ranking weapons. For example, at the outset with everything being equal in a number of weapons, one would select the weapon with the smallest P_k variance. Finally, with an objective towards reducing the P_k variance, the methodology developed allows one to breakdown the P_k variance into its components and to identify those input parameters with the largest variance contribution. This then can form a first step towards developing appropriate variance reduction techniques to minimize the error in P_k.

In what follows, if Z is a random variable, then $\bar{Z} = E[Z]$ refers to the expectation of Z. $\mathrm{Var}[Z]$ refers to the variance of Z and σ_Z refers to the standard deviation of Z ($\sigma_Z^2 = \mathrm{Var}[Z]$).

In cases involving aiming error, such error is assumed to be unbiased, that is to be centered at the location of the point target and to be represented by Gaussian (normal) distributions in each of the directions of the flightpath of the airplane (the range direction or the x - axis) and perpendicular to such flightpath (the deflection direction or the y - axis). The errors in each of these directions are assumed to be independent with respective standard deviations σ_x and σ_y. All targets are assumed to be surface targets.

II. OBJECTIVES

The main objective of this project is to develop a methodology for computing confidence intervals for the probability of kill. We do not look at other weapon effectiveness measure such as the mean area of effectiveness (MAE) or the fractional coverage. Also, we do not attempt to apply the methodology to the general problem involving an arbitrary weapon-

target situation. Instead, we treat some simple, but practical, situations leaving the more complicated cases for later investigations. Our specific objectives are:

(1) To develop several methods for estimating the mean and the variance of P_k for some specific situations. Although three methods are provided, nevertheless, only the first two methods are extensively used, the third one being quite restrictive in its application.

(2) To arrive at closed form expressions for $E[P_k]$ and Var $[P_k]$ and to compare through the use of numerical examples the various methods.

(3) To apply the methodologies to the four following situations:

   a. Fragment sensitive target in the presence of no aiming error.

   b. Fragment sensitive targets in the presence of aiming error.

   c. Blast sensitive targets in the presence of no aiming error.

   d. Blast sensitive targets in the presence of aiming errors.

In establishing confidence intervals, a two-standard deviation, two-sided confidence interval is selected to provide an interval estimation to the probability of kill. In such a case, Chebyshev's inequality guarantees at least a 75 percent confidence interval since for $\alpha = 2$, we have

$$P\left\{ E[P_k] - \alpha \sigma_{P_k} \leq P_k \leq E[P_k] + \alpha \sigma_{P_k} \right\} \geq 1 - \frac{1}{\alpha^2} = .75$$

III. METHODOLOGY

For argument sake, suppose that the probability of kill, P_k, is a function of the two input parameters X and Y so that $P_k = P_k(X, Y)$. Suppose that each of the parameters X and Y are estimates subject to error. It is required to determine $E[P_k]$ and Var $[P_k]$. We digress on three approaches that might be used to obtain the required result:

Method 1

If the joint distribution of X and Y is known and is say $f_{X,Y}(x, y)$, then

$$E[P_k] = \iint_{x,y} P_k(x,y)\, f_{X,Y}(x,y)\, dx\, dy$$

and

$$E\left[P_k^2\right] = \iint_{x,y} P_k^2(x,y)\, f_{X,Y}(x,y)\, dx\, dy$$

From these two expressions, one can obtain

$$Var\left[P_k\right] = E\left[P_k^2\right] - E^2\left[P_k\right]$$

In many practical situations, it is just not possible to completely characterize the function $f_{X,Y}(x, y)$. Further, very often, subjective judgements enter in the statistical estimation of X and Y. Under these conditions, can one still obtain values for $E\left[P_k\right]$ and $Var\left[P_k\right]$? The answer, of course, depends on how acccurately these values should be measured. For the particular applications contemplated, a precise measurement was not a major objective; rather, a reasonable approximation was expected. To this end, it was assumed that X and Y were independent random variables uniformly distributed over the respective ranges $x_1 < x < x_2$ and $y_1 < y < y_2$. The two major advantages of this approach are the following:

(1) It proves helpful in using subjective measurements to the problem; for then, a minimum value and a maximum value of X and Y can be easily specified without any further additional information. Note here that under these assumptions, the mean and variance of X and Y are given respectively by

$$E[X] = \frac{x_1 + x_2}{2} \quad ; \quad Var[X] = \frac{(x_2 - x_1)^2}{12}$$

$$E[Y] = \frac{y_1 + y_2}{2} \quad ; \quad Var[Y] = \frac{(y_2 - y_1)^2}{12}$$

(2) It often provides expressions for $E\left[P_k\right]$ and $E\left[P_k^2\right]$ which can be analytically manipulated to arrive at closed form expressions.

Method 2

Given $P_k(X, Y)$ to be a continuous and differentiable function of X and Y, then as a first approximation

$$E\left[P_k(X, Y)\right] = P_k(\bar{X}, \bar{Y}) \qquad (1)$$

One however has to be cautious in utilizing this approach for it is possible to give counterexamples for which this approximation totally fails to hold. However, in all the situations that we will deal with for which

Method 1 proves applicable, the use of (1) will show consistently a close accuracy to Method 1.

If one expands P_k (X, Y) as a Taylor's series about the point $(\bar{X}, \bar{Y})$, then to a first approximation

$$P_k(X,Y) = P_k(\bar{X},\bar{Y}) + (X-\bar{X})\left.\frac{\partial P_k}{\partial X}\right|_{\bar{X},\bar{Y}} + (Y-\bar{Y})\left.\frac{\partial P_k}{\partial Y}\right|_{\bar{X},\bar{Y}} \qquad (2)$$

If X and Y are independently distributed, then taking variances on both sides of (2) yields

$$\mathrm{Var}[P_k] = \left\{\left.\frac{\partial P_k}{\partial X}\right|_{\bar{X},\bar{Y}}\right\}^2 \mathrm{Var}[X] + \left\{\left.\frac{\partial P_k}{\partial Y}\right|_{\bar{X},\bar{Y}}\right\}^2 \mathrm{Var}[Y] \qquad (3)$$

As can be noted, the main advantage in using formula (3) is that it provides a way for segregating the contribution of each variance component to the total variance. An expression similar to (3) can be obtained if X and Y are correlated.

Method 3

This method appears to have originated in statistical economics in the study of index numbers. Although this method will not be used in the sequel, nevertheless, it merits to be mentioned as a possible approach for solving simply structured problems. The method is best illustrated through the use of a specific example. Let $P_k = X/Y$; write

$$X = \bar{X} + U \qquad \text{and} \qquad Y = \bar{Y} + U$$

where U and V are random variables with zero mean such that

$$\mathrm{Var}[X] = \mathrm{Var}[U] = \sigma_X^2 \quad ; \quad \mathrm{Var}[Y] = \mathrm{Var}[V] = \sigma_Y^2$$

Then

$$P_k = \frac{\bar{X}}{\bar{Y}}\left(1+\frac{U}{\bar{X}}\right)\left(1+\frac{V}{\bar{Y}}\right)^{-1}$$

Expanding the last factor by the binomial theorem and retaining the first few terms, we obtain

$$P_k \simeq \frac{\bar{X}}{\bar{Y}}\left(1+\frac{U}{\bar{X}}\right)\left(1-\frac{V}{\bar{Y}}+\frac{V^2}{\bar{Y}^2}\right)$$

$$\simeq \frac{\bar{X}}{\bar{Y}}\left(1+\frac{U}{\bar{X}}-\frac{V}{\bar{Y}}-\frac{UV}{\bar{X}\bar{Y}}+\frac{V^2}{\bar{Y}^2}\right)$$

Taking expectations on both sides of this relation yields:

$$E[P_k] \simeq \frac{\bar{X}}{\bar{Y}} \left\{ 1 - \frac{E[UV]}{\bar{X}\bar{Y}} + \frac{E[V^2]}{\bar{Y}^2} \right\}$$

$$\simeq \frac{\bar{X}}{\bar{Y}} \left(1 - \frac{\rho \sigma_X \sigma_Y}{\bar{X}\bar{Y}} + \frac{\sigma_Y^2}{\bar{Y}^2} \right)$$

Where ρ is the correlation coefficient between U and V.
A similar approach may be used to obtain $E[P_k^2]$ resulting in

$$E[P_k^2] \simeq \frac{\bar{X}}{\bar{Y}} \left(1 + \frac{\sigma_X^2}{\bar{X}^2} + 3\frac{\sigma_Y^2}{\bar{Y}^2} - 4\frac{\rho \sigma_X \sigma_Y}{\bar{X}\bar{Y}} \right)$$

Var $[P_k]$ is immediately obtained from

$$Var[P_k] = E[P_k^2] - E^2[P_k]$$

IV. FRAGMENT SENSITIVE TARGETS

We have seen that for fragment sensitive targets, given that the weapon bursts at (0, 0), the probability of kill P_{kf} for a target point situated at (x, y) is

$$P_{kf} = \exp\left[-\left(\frac{x^2}{R_x^2} + \frac{y^2}{R_y^2}\right)\right] \tag{4}$$

where R_x and R_y are the two weapon radii, and represent the two input parameters. We consider now the two cases when aiming error is either absent or present.

A. No Aiming Error Present

Method 1

Assume R_x and R_y to be independently and uniformly distributed over the respective ranges $R_{x_1} < R_x < R_{x_2}$ and $R_{y_1} < R_y < R_{y_2}$.
Then we have

$$E[P_{kf}] = \frac{1}{(R_{y_2} - R_{y_1})(R_{x_2} - R_{x_1})} \int_{R_{y_1}}^{R_{y_2}} \int_{R_{x_1}}^{R_{x_2}} e^{-\left(\frac{x^2}{R_x^2} + \frac{y^2}{R_y^2}\right)} dR_x\, dR_y \tag{5}$$

Let

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{w^2}{2}}\, dw \tag{6}$$

It can be shown that $E[P_{kf}]$ is given by the following expression

$$E[P_{kf}] = \frac{1}{(R_{y_2} - R_{y_1})(R_{x_2} - R_{x_1})}$$

$$\cdot \left\{ R_{x_2}\, e^{-\frac{x^2}{R^2_{x_2}}} - R_{x_1}\, e^{-\frac{x^2}{R^2_{x_1}}} - 2x\sqrt{\pi}\left[\Phi\left(\frac{x\sqrt{2}}{R_{x_1}}\right) - \Phi\left(\frac{x\sqrt{2}}{R_{x_2}}\right)\right]\right\}$$

$$\cdot \left\{ R_{y_2}\, e^{-\frac{y^2}{R^2_{y_2}}} - R_{y_1}\, e^{-\frac{y^2}{R^2_{y_1}}} - 2y\sqrt{\pi}\left[\Phi\left(\frac{y\sqrt{2}}{R_{y_1}}\right) - \Phi\left(\frac{y\sqrt{2}}{R_{y_2}}\right)\right]\right\} \quad (7)$$

The expression for $E[P^2_{kf}]$ is obtained from the previous expression by by substituting $x\sqrt{2}$ for x and $y\sqrt{2}$ for y.

Method 2

The results of using this method are:

$$E[P_{kf}] = \exp\left[-\left(\frac{x^2}{R^2_x} + \frac{y^2}{R^2_y}\right)\right] \quad (8)$$

$$Var[P_{kf}] = 4\exp\left[-2\left(\frac{x^2}{R^2_x} + \frac{y^2}{R^2_y}\right)\right]\left\{\frac{x^4}{R^4_x}\, Var[R_x] + \frac{y^4}{R^4_y}\, Var[R_y]\right\} \quad (9)$$

Numerical Example

Expressing all linear units in feet, let

$R_x = 80$; $R_x = 90$; $E[R_x] = 85$; $Var[R_x] = 100/12$

$R_y = 160$; $R_y = 180$; $E[R_y] = 170$; $Var[R_y] = 400/12$

$x = 50$; $y = 100$

$$P_{kf} = E[P_{kf}] \pm 2\sigma_{P_{kf}}$$

is given by $P_{kf} = .4996 \pm .0356$ (Method 1)

and $P_{kf} = .500 \pm .034$ (Method 2)

B. Aiming Error Present

Assume that the x - axis to be oriented in the direction of range and the y - axis to be oriented in the direction of deflection. The target is located at (x, y) and the center of weapon impact is (u, v). The probability that the weapon will impact in an interval (du, dv) in the neighborhood of the point (u, v) is

$$g(u-x, v-y)\,du\,dv = \frac{1}{2\pi\sigma_x\sigma_y}\, e^{-\left[\frac{(u-x)^2}{2\sigma_x^2} + \frac{(v-y)^2}{2\sigma_y^2}\right]}\, du\,dv \qquad (10)$$

The probability of kill due to fragmentation at the point (x, y) given that the weapon bursts at (u, v) is given by the Carleton damage function

$$P(x-u, y-v) = e^{-\left[\frac{(x-u)^2}{R_x^2} + \frac{(y-v)^2}{R_y^2}\right]} \qquad (11)$$

The probability of kill at (x, y) due to fragmentation is

$$P_{kf} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\text{Probability of kill at } (x, y) \,\middle|\, \text{weapon impact at } (u, v)\right] \cdot \left[\text{Probability of weapon impact at } (u, v)\right] du\,dv$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x-u, y-v)\, g(u-x, v-y)\, du\,dv$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\left[\frac{(x-u)^2}{R_x^2} + \frac{(y-v)^2}{R_y^2}\right]} \cdot \frac{1}{2\pi\sigma_x\sigma_y}\, e^{-\left[\frac{(u-x)^2}{2\sigma_x^2} + \frac{(v-y)^2}{2\sigma_y^2}\right]} du\,dv \qquad (12)$$

This derivation of P_{kf} differs from [1] in that it uses a more general approach based on conditional probabilities. Making the change in variables

$$\frac{\theta_1}{\sqrt{2}} = \sqrt{\frac{1}{R_x^2} + \frac{1}{2\sigma_x^2}}\,(u-x) \quad ; \quad \frac{\theta_2}{\sqrt{2}} = \sqrt{\frac{1}{R_y^2} + \frac{1}{2\sigma_y^2}}\,(v-y)$$

we obtain $P_{kf} = I_x I_y$

where

$$I_x = \frac{R_x}{\sqrt{R_x^2 + 2\sigma_x^2}} \quad ; \quad I_y = \frac{R_y}{\sqrt{R_y^2 + 2\sigma_y^2}}$$

Here

$$E[P_{kf}] = E[I_x] \cdot E[I_y]$$

$$Var[P_{kf}] = E[I_x^2]\, E[I_y^2] - \left\{ E[I_x] \cdot E[I_y] \right\}^2$$

This holds under the assumption that R_x, R_y, σ_x and σ_y are independently distributed.

Method 1

Assume R_x, R_y, σ_x and σ_y to be independently and uniformly distributed over the respective ranges

$$R_{x_1} < R_x < R_{x_2} \quad ; \quad R_{y_1} < R_y < R_{y_2}$$
$$\sigma_{x_1} < \sigma_x < \sigma_{x_2} \quad ; \quad \sigma_{y_1} < \sigma_y < \sigma_{y_2}$$

Let

$$X_{ij} = \sqrt{1 + \frac{R_{x_i}^2}{2\sigma_{x_j}^2}} \quad ; \quad Y_{ij} = \sqrt{1 + \frac{R_{y_i}^2}{2\sigma_{y_j}^2}} \qquad (i,j = 1,2)$$

The results for $E[I_x]$ and $E[I_x^2]$ are as follows:

$$E[I_x] = \frac{1}{\sqrt{2}\,(R_{x_2} - R_{x_1})(\sigma_{x_2} - \sigma_{x_1})} \left[\sigma_{x_2}^2 (X_{22} - X_{12}) - \sigma_{x_1}^2 (X_{21} - X_{11}) \right.$$

$$\left. + \frac{R_{x_2}^2}{2} \ln \frac{\sigma_{x_2}(1 + X_{22})}{\sigma_{x_1}(1 + X_{21})} - \frac{R_{x_1}^2}{2} \ln \frac{\sigma_{x_2}(1 + X_{12})}{\sigma_{x_1}(1 + X_{11})} \right] \qquad (13)$$

$$E[I_x^2] = \frac{1}{2} + \frac{\sqrt{2}}{(R_{x_2} - R_{x_1})(\sigma_{x_2} - \sigma_{x_1})} \left[\frac{R_{x_2}^2 + 2\sigma_{x_2}^2}{4} \tan^{-1} \frac{\sigma_{x_2}\sqrt{2}}{R_{x_2}} \right.$$

$$- \frac{R_{x_2}^2 + 2\sigma_{x_1}^2}{4} \tan^{-1} \frac{\sigma_{x_1}\sqrt{2}}{R_{x_2}} - \frac{R_{x_1}^2 + 2\sigma_{x_2}^2}{4} \tan^{-1} \frac{\sigma_{x_2}\sqrt{2}}{R_{x_1}}$$

$$+ \frac{R_{x_1}^2 + 2\sigma_{x_1}^2}{4} \tan^{-1} \frac{\sigma_{x_1}\sqrt{2}}{R_{x_1}} \Bigg] \qquad (14)$$

The expressions for $E[I_y]$ and $E[I_y^2]$ are easily obtained from the previous expressions by substituting y for x and Y for X.

Method 2

The results of method 2 are

$$E[P_{kf}] = \frac{\bar{R}_x}{\sqrt{\bar{R}_x^2 + 2\bar{\sigma}_x^2}} \cdot \frac{\bar{R}_y}{\sqrt{\bar{R}_y^2 + 2\bar{\sigma}_y^2}} \qquad (15)$$

$$\mathrm{Var}[P_{kf}] = \frac{4\bar{\sigma}_x^2 \bar{R}_y^2}{(\bar{R}_x^2 + 2\bar{\sigma}_x^2)^3 (\bar{R}_y^2 + 2\bar{\sigma}_y^2)} \left\{ \bar{\sigma}_x^2 \mathrm{Var}[R_x] + \bar{R}_x^2 \mathrm{Var}[\sigma_x] \right\}$$

$$+ \frac{4\bar{\sigma}_y^2 \bar{R}_x^2}{(\bar{R}_x^2 + 2\bar{\sigma}_x^2)(\bar{R}_y^2 + 2\bar{\sigma}_y^2)^3} \left\{ \bar{\sigma}_y^2 \mathrm{Var}[R_y] + \bar{R}_y^2 \mathrm{Var}[\sigma_y] \right\} \qquad (16)$$

Numerical Example

Expressing all linear units in feet, let

$R_{x_1} = 80$; $R_{x_2} = 90$; $\bar{R}_x = 85$; $\mathrm{Var}[R_x] = 100/12$

$R_{y_1} = 162$; $R_{y_2} = 178$; $\bar{R}_y = 170$; $\mathrm{Var}[R_y] = 256/12$

$\sigma_{x_1} = 100$; $\sigma_{x_2} = 200$; $\bar{\sigma}_x = 150$; $\mathrm{Var}[\sigma_x] = 10{,}000/12$

$\sigma_{y_1} = 70$; $\sigma_{y_2} = 90$; $\bar{\sigma}_y = 80$; $\mathrm{Var}[\sigma_y] = 400/12$

$$P_{kf} = E P_{kf} \pm 2\sigma_{P_{kf}}$$

is given by $P_{kf} = .318 \pm .110$ (Method 1)

and $P_{kf} = .310 \pm .106$ (Method 2)

V. BLAST SENSITIVE TARGETS

Let the blast occur at an origin point, and let R be the distance from the center of the blast to a point target. Then the probability of kill

due to blast P_{kb} is

$$P_{kb} = \begin{cases} 1 & 0 < R < A \\ 1 - \frac{R-A}{B-A} & A \le R < B \\ 0 & B \le R < \infty \end{cases} \qquad (17)$$

where A is the maximum distance at which irreparable damage is experienced and B is the minimum distance at which no damage occurs.

A. No Aiming Error Present

Method 1

Assume A and B are independently and uniformly distributed over the ranges $A_1 < A < A_2$ and $B_1 < B < B_2$. $(A_1 < A_2 < B_1 < B_2)$. One has to consider different situations depending on the relative positions of the points A and B. We illustrate this by working out a given situation. Assume $A_1 < R < A_2$, then

if $A_1 < R < A < A_2$, $P_{kb} = 1$, and

if $A_1 < A < R < A_2$, $P_{kb} = 1 - \frac{R-A}{B-A}$

Then

$$E[P_{kb}] = \int_{B_1}^{B_2} \left\{ \int_{A_1}^{R} \left[1 - \frac{R-A}{B-A} \right] \frac{1}{A_2 - A_1} dA + \int_{R}^{A_2} 1 \cdot \frac{1}{A_2 - A_1} dA \right\} \frac{1}{B_2 - B_1} dB$$

$$E[P_{kb}^2] = \int_{B_1}^{B_2} \left\{ \int_{A_1}^{R} \left[1 - \frac{R-A}{B-A} \right]^2 \frac{1}{A_2 - A_1} dA + \int_{R}^{A_2} 1^2 \cdot \frac{1}{A_2 - A_1} dA \right\} \frac{1}{B_2 - B_1} dB$$

The other possibilities are worked out using a similar approach.

Method 2

This method is not applicable since Pkb is not a differentiable function of the parameters A and B.

Numerical Example

Let $A_1 = 13$, $A_2 = 17$, $B_1 = 20$, $B_2 = 22$, $R = 19$ (all in feet)

$$P_{kb} = E[P_{kb}] \pm 2\sigma_{P_{kb}}$$

is given by $P_{kb} = .3399 \pm .1814$ (Method 1)

B. Aiming Error Present

Let the target be located at (x, y). Due to aiming error the weapon bursts in an interval (du, dv) in the neighborhood of (u, v) and with probability

$$g(u - x, v - y)\, du\, dv = \frac{1}{2\pi \sigma_x \sigma_y} e^{-\left[\frac{(u-x)^2}{2\sigma_x^2} + \frac{(v-y)^2}{2\sigma_y^2} \right]} du\, dv \quad (18)$$

The probability of a kill at (x, y) due to blast given that the point of impact is (u, v) is assumed to depend on the distance between (x, y) and (u, v) or $r = \sqrt{(x - u)^2 + (y - v)^2}$ and is given by

$$\bar{P}_{kb}(r) = \begin{cases} 1 & 0 < r < A \\ 1 - \frac{r - A}{B - A} & A \le r < B \\ 0 & B \le r < \infty \end{cases} \quad (19)$$

The probability of kill due to blast is thus

$$P_{kb} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \bar{P}_{kb} \left[\sqrt{(x-u)^2 + (y-v)^2} \right] g(u-x, v-y)\, du\, dv$$

Let $u - x = r \cos\theta$, $v - y = r \sin\theta$. Then

$$P_{kb} = \frac{1}{2\pi \sigma_x \sigma_y} \int_0^{2\pi} \int_0^{A} e^{-\frac{r^2 \cos^2\theta}{2\sigma_x^2} - \frac{r^2 \sin^2\theta}{2\sigma_y^2}} r\, dr\, d\theta$$

$$+ \frac{1}{2\pi \sigma_x \sigma_y} \int_0^{2\pi} \int_A^{B} \left[1 - \frac{r - A}{B - A} \right] e^{-\frac{r^2 \cos^2\theta}{2\sigma_x^2} - \frac{r^2 \sin^2\theta}{2\sigma_y^2}} r\, dr\, d\theta \quad (20)$$

Let now

$$a = \frac{1}{4} \left(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_y^2} \right) \quad ; \quad b = \frac{1}{4} \left(\frac{1}{\sigma_y^2} - \frac{1}{\sigma_x^2} \right)$$

Then, it can be shown that

$$P_{kl} = \frac{1}{\sigma_x \sigma_y} \int_0^A r e^{-ar^2} I_0(br^2)\, dr + \frac{1}{\sigma_x \sigma_y} \int_A^B \left(\frac{B}{B-A} - \frac{r}{B-A} \right) r e^{-ar^2} I_0(br^2)\, dr \qquad (21)$$

where $I_0(.)$ is the modified Bessel function of the first kind of zero order.

When $\sigma_x = \sigma_y = \sigma$, it can be shown that

$$P_{kl} = 1 - \frac{\sqrt{2\pi}\sigma}{(B-A)} \left[\Phi\left(\frac{B}{\sigma}\right) - \Phi\left(\frac{A}{\sigma}\right) \right] \qquad (22)$$

Method 1

It was not possible to obtain closed form expressions for any of the above cases using method 1. However, when $\sigma_x = \sigma_y = \sigma$ and $\frac{B}{\sigma} \ll 1$, approximations for $E[P_{kb}]$ and $Var[P_{kb}]$ can be obtained.

Method 2

The use of method 2 in the general case when $\sigma_x \neq \sigma_y$ gives rise to integrals involving Bessel functions which cannot be expressed in closed forms. However, when $\sigma_x = \sigma_y = \sigma$, the use of method 2 provides expressions involving the normal probability function. Let

$$v(\bar{A}, \bar{B}, \bar{\sigma}) = \left[\Phi\left(\frac{\bar{B}}{\bar{\sigma}}\right) - \Phi\left(\frac{\bar{A}}{\bar{\sigma}}\right) \right] \frac{\sqrt{2\pi}}{(\bar{B} - \bar{A})} \qquad (23)$$

Then,

$$E[P_{kl}] = 1 - \bar{\sigma}\, v(\bar{A}, \bar{B}, \bar{\sigma}) \qquad (24)$$

and

$$Var[P_{kl}] = \left[\frac{1}{(\bar{B} - \bar{A})\bar{\sigma}} \left(\bar{A} e^{-\frac{\bar{A}^2}{2\bar{\sigma}^2}} - \bar{B} e^{-\frac{\bar{B}^2}{2\bar{\sigma}^2}} \right) + v(\bar{A}, \bar{B}, \bar{\sigma}) \right]^2 Var[\sigma] \qquad (25)$$

$$+\left[-\frac{\bar{\sigma}}{(\bar{B}-\bar{A})}\,U(\bar{A},\bar{B},\bar{\sigma})+\frac{1}{(\bar{B}-\bar{A})}\,e^{-\frac{\bar{A}^2}{2\bar{\sigma}^2}}\right]^2 \mathrm{Var}[A]$$

$$+\left[\frac{\bar{\sigma}}{(\bar{B}-\bar{A})}\,U(\bar{A},\bar{B},\bar{\sigma})+\frac{1}{(\bar{B}-\bar{A})}\,e^{-\frac{\bar{B}^2}{2\bar{\sigma}^2}}\right]^2 \mathrm{Var}[B] \qquad (26)$$

Numerical Example

Assuming all units are in feet, let

$A_1 = 13$; $A_2 = 17$; $\bar{A} = 15$; Var $[A] = 16/12$

$B_1 = 20$; $B_2 = 22$; $\bar{B} = 21$; Var $[B] = 4/12$

$\sigma_1 = 40$, $\sigma_2 = 60$; $\bar{\sigma} = 50$; Var $[\sigma] = 400/12$

$$P_{kb} = E[P_{kb}] \pm 2\,\sigma_{P_{kb}}$$

is given by $P_{kb} = .0682 \pm .0236$ (Method 1)

and $P_{kb} = .0632 \pm .0294$ (Method 2)

VI. RECOMMENDATIONS

The methodology developed in this report for measuring $E[P_k]$ and Var $[P_k]$ provides the foundation for solving more complex problems in the area of weapon systems analysis. In practice, it is extremely difficult to determine the probability distribution function of the parameters entering in the evaluation of P_k. Determination of their first two moments is more in the realm of reality. Therefore, it seems logical that any estimation of P_k relies only on the scant information that one is dealing with in practice. This brings up the next problem that should be investigated in this research area: that of accuracy of estimation. Is it possible to ascertain that with the methods used one can provide for a given confidence interval, confidence coefficients better than the minimum dictated by the Chebyshev's inequality? At least for Method 1 (in which the assumption of uniform distribution is made) the answer is in the positive since the true distribution function of P_k can be theoretically derived. However, this may not be practical and one may have to resort to other methodologies.

We have considered so far, the two cases when the basic statistics (mean and standard deviation) of the probability of kill due to blast (P_{kb}) and due to fragmentation (P_{kf}) were separately estimated. In most practical problems, these two after-effects of an exploding weapon are

present simultaneously and the kill effect of the weapon may be the result of either blast or fragmentation or both. It may be assumed that kill due to blast and kill due to fragmentation are either two dependent or two independent events. In as much as the estimation problem is concerned, both P_{kb} and P_{kf} are random variables subject to estimation error. Therefore, another recommendation for follow-on research using the methods of the present work is to treat the more general case when P_{kb} and P_{kf} are simultaneously present. It is unlikely that general closed form expressions will result from such an investigation, and one would logically rely on numerical techniques to obtain the necessary values for $E[P_k]$ and $\text{Var}[P_k]$.

In addition to the probability of kill, the lethal area and the fractional coverage are also used as measures to evaluate weapon effectiveness under prescribed situations. A closely related topic involves the use of damage functions other than the ones considered so far, such as the three-parameter Carleton damage functions for fragment-sensitive targets and the cookie-cutter functions for impact sensitive targets. As a consequence, a further recommendation is to extend the methodology to these other areas which were not treated in this report.

When evaluating $E[P_{kf}]$ and $\text{Var}[P_{kf}]$, the Carleton damage function was expressed in terms of the two weapon radii. In many situations, the means area of effectiveness (MAE) as well as the weapon impact angle are entered as inputs. The prescribed method could be adapted for this particular situation and measures of $E[P_{kf}]$ and $\text{Var}[P_{kf}]$ would then be expressible directly in terms of the given input parameters.

We have also seen that when computing probability of kill due to blast, P_{kb}, the general case involving different aiming errors in range and in deflection resulted in integrals of modified Bessel function of zero order, which could not be evaluated in closed forms. To further this study, it is recommended that approximation techniques, such as asymptotic expansions, be used to obtain expressions for $E[P_{kb}]$ and $\text{Var}[P_{kb}]$.

In performing vulnerability analysis, the fragmentation data are used to determine the MAE of a weapon as well as the parameters of the Carleton damage function. A further recommendation is to analyze statistically these field data to provide the necessary information to numerically estimate the mean and the variance of the various input parameters.

As a further extension, it was pointed out to the present author that the methodology could be applied to the general problem in which the probability of kill P_k is expressible as a monotone function of the distance R between target and point of weapon firing. The target is killed either through direct hit or through blast or through fragmentation. It appears that little work has been done in the theoretical derivation of the functional relationship between P_k and R, knowing the characteristics of the weapon and the target.

Finally, the methodology described could be used for the following cases:

a. Area targets rather than point targets
b. Cluster weapons
c. Guided weapons
d. Laser-guided weapons
e. Stick delivery of weapons
f. Runways targets
g. Biased aiming of weapons
h. Presence of ballistic error

REFERENCES

1. "Derivation of JMEM/AS Open-End Methods", Joint Technical Coordinating Group for Munitions Effectiveness, 61 JTCG/ME-3-7, May 1980
2. D. G. DiCanio, "Error Analysis - Lethality and Safe Separation Parameters for Air-Delivered Fragmenting Munitions, (Summary Report of Phase 1)," Technical Report AFATL-TR-67-183, Air Force Armament Laboratory, A F Systems Command, Eglin AFB FL, October 1967
3. K. A. Myers, "Lethal Area Description," Technical Note No. 1510, Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland, July 1963 (Available from DDC, DSA)
4. F. A. Roescher, "Measures of Effectiveness for Conventional Munitions," Report No. 2350-72, Air University, Maxwell AFB, Alabama, May 1972
5. G. Schroeter, "The Variance of the Coverage of a Randomly Located Area Target by a Salvo of Weapons," Naval Research Logistics Quarterly, Vol 29, No. 1, March 1982
6. R. Snow, "Fast-Val: A Theoretical Approach to Some General Target Coverage Problems," Memorandum RM-4566-PR, The Rand Corporation, Santa Monica, California, March 1966
7. R. Snow and M. Ryan, "A Simplified Weapons Evaluation Model," Memorandum RM-5677-1-PR, The Rand Corporation, Santa Monica, California, May 1970

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

METHODS FOR COMPUTATION OF CERTAIN DIAGNOSTICS FOR THE AFGL GENERAL CIRCULATION MODEL

Prepared by:	Dr. Stanley L. Spiegel
Academic Rank:	Assistant Professor
Department and University:	Mathematics Department University of Lowell
Research Location:	Air Force Geophysics Laboratory Meteorology Division/Atmospheric Prediction Branch
USAF Research Colleague:	Dr. Chien-hsiung Yang

Date:	August 25, 1982
Contract No:	F49620-82-C-0035

METHODS FOR COMPUTATION OF CERTAIN DIAGNOSTICS FOR THE AFGL GENERAL CIRCULATION MODEL

by

Stanley L. Spiegel

ABSTRACT

Computer programs have been developed to permit the calculation of kinetic and potential energy, and enstrophy for the atmospheric motion fields associated with the primitive equation spectral general circulation model under development at the Air Force Geophysics Laboratory. These calculations can be performed in either of the two physical space coordinate systems relevant for assessing model performance. Results of these calculations, and similar computations being carried out in the model's spectral domain, will be helpful in understanding and improving the functioning of the model.

Acknowledgement

It is a pleasure to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for making possible my appointment to the Air Force Geophysics Laboratory under the Summer Faculty Research Program. I am grateful to the Atmospheric Prediction Branch and its Chief, Mr. Donald Chisholm, for welcoming me and affording me a most pleasant environment in which to work.

I thank Dr. Chien-hsiung Yang for suggesting this interesting problem to me and for many informative and stimulating discussions regarding my progress during the summer. I also benefited greatly from numerous technical discussions with Dr. Stephen Brenner, Mr. Samuel Yee and Dr. Ken Mitchell. Lastly I am appreciative of the thoughtful programming assistance provided by Mr. Karekin Agazarian and of the cheerful and skillful secretarial assistance of Mrs. Betty Blanchard.

I. INTRODUCTION

The Atmospheric Prediction Branch of the Meteorology Division, Air Force Geophysics Laboratory, is in the process of developing a general circulation model patterned after the primitive equation spectral model employed at the National Meteorological Center[1] (NOAA). It is of course necessary to evaluate the performance of this model to ensure that essential physical processes are correctly incorporated into a suitable numerical computation scheme. One important test is to compare fields predicted by the model, such as the surface pressure or 500mb geopotential, with those actually observed. Objective methods for evaluating such comparison (so called skill scores) are discussed in the literature,[2,3] and the required software for performing this sort of analysis already exists. The focus of this summer's work has been to develop procedures for an additional method of assessing model performance. This involves the computation of certain model diagnostics - kinetic and potential energy, and enstrophy - in order to compare predicted values with those observed in the atmosphere.

In order to understand how these model diagnostics should be computed, it is necessary to have an overview of the functioning of the model. The input data (so called FGGE data, for First GARP Global Experiment) is provided at a regularly spaced latitude-longitude array of grid points (2.5° spacing between points in both directions) and at 12 mandatory pressure levels (1000, 850, 700, 500, 400, 300, 250, 200, 150, 100, 70, and 50mb) in the vertical. This data consists of geopotential height, velocity and temperature. (Surface values of temperature and geopotential are also given; surface velocities and

pressures are readily obtainable via extrapolation procedures from the given values.) We shall refer to this grid system as the "FGEE" grid.

The model however requires that its fields be given in a different physical grid system. This grid has uniform spacing in longitude, but latitudinal values are given at the latitudes required in order to utilize Gaussian quadrature in the latitudinal direction. The vertical coordinate is not pressure, but rather the sigma coordinate introduced by Phillips[4]. In its present form, this "model" grid has a longitudinal spacing of 7.5°, 40 Gaussian latitudes pole-to-pole, and six sigma levels in the vertical, bounded by sigma values 1.0, .90, .75, .50, .25, .15, and 0.0. The field values in the model grid are obtained from the FGGE grid values in a so called pre-processing phase which consists principally of interpolation procedures to compute the required values.

Then after an initialization procedure to remove undesirable modes from the input data, the fields are expressed as truncated series of surface spherical harmonics (in each σ-level), and the predictive equations are solved spectrally[1,5]. From the predicted spectral components, the new fields can be constructed in physical space at the model grid points and, after a post-processing procedure, the model forecasts are available on the FGGE grid.

Since error can be introduced during the pre- and post-processing procedures and in the initialization procedure, as well as in the prediction process, it is clearly desirable to be able to compute diagnostics in terms of FGGE grid values, model grid values, and spectral values. This would help distinguish errors in model prediction from those caused by the transference of values from one representation to another, and thus suggest relevant areas for model improvement.

II. OBJECTIVES

The main objective of this project has been to develop computer programs for the computation of model diagnostics in the two physical space coordinate systems - the FGGE and model grids - utilized by the primitive equation spectral model. (Computations in the spectral representation of the model of total energy had already been performed, and the remainder of the spectral diagnostic calculations have been programmed, by Dr. Stephen Brenner of AFGL.) These programs were to be written in such a way that level by level subtotals of these diagnostics would be available for analysis. It was required that the algorithms be consistent with the core capacity of the CDC 6600 computer now in use at AFGL (about 100K words) and be easily transportable for implementation on other computers.

A further goal was to compute the global available potential energy; this was in fact accomplished only for the FGGE grid system.

III. FGGE GRID DIAGNOSTICS

The FGGE grid system consists of a three-dimensional array of points: 144 in the west-to-east direction at 2.5° longitudinal increments (indexed by I), 73 in the north-to-south direction at 2.5° latitudinal spacing (indexed by J), and 12 in the vertical at the mandatory pressure levels (indexed by K). This gives a total of 144x73x12 or 126,144 points for each field represented on this grid. (The values at the north and south poles (J=1 and 73 respectively) are given redundantly - 144 times each - at each K level; this feature has certain computational advantages in other contexts). The index of the pressure levels increases with height, such that the 1000mb level is denoted by a K

value of 2 and the uppermost 50mb level by a K value of 13; the reason for using these values rather than 1 to 12 will become apparently shortly.

The quantities to be computed are the following, in discreet representation:

Kinetic energy: $$\frac{1}{2}\sum\left[U^2(I,J,K)+V^2(I,J,K)\right]\cdot W(I,J,K) \qquad (1)$$

Gravitational potential energy: $$\frac{1}{g}\sum\Phi(I,J,K)\cdot W(I,J,K) \qquad (2)$$

Internal energy: $$\frac{C_v}{g}\sum T(I,J,K)\cdot W(I,J,K) \qquad (3)$$

Enstrophy: $$\frac{1}{2}\sum\zeta^2(I,J,K)\cdot W(I,J,K) \qquad (4)$$

where the sums extend over all I and J at K values above the earth's surface and at surface points as well. Here U and V represent zonal and meridional velocity components, Φ is geopotential height, T is temperature, ζ is the local vertical component of vorticity computed on isobaric surfaces, Cv is the specific heat of air at constant volume ($717 JK^{-1}kg^{-1}$) and g is gravitational acceleration (9.81 $cm\text{-}sec^{-2}$). The weights W represent the mass of atmosphere represented by the grid point denoted I, J, K. Given the size of the arrays we are dealing with, the computations are necessarily performed in stages, one level at a time. First the weight field is computed, and stored. Then for each desired diagnostic, relevant data fields are read in for at a given K level so that the diagnostic quantity can be calculated at each point of that level. Next the weights for that level are read in so that the product of diagnostic values times weights can be summed producing the desired diagnostic subtotal for the given level. When

this has been done for each K level, the global value (grand total) of the diagnostic can be calculated.

The central task is the determination of the appropriate weights. That is, we seek a suitable discreet approximation of an element of atmospheric mass

$$dm = a^2 \rho \cos\phi \, d\phi \, d\lambda \, dz \qquad (5)$$
$$= (a^2/g) \cos\phi \, d\phi \, d\lambda \, dP$$

where ρ is the atmospheric density, λ is longitude, ϕ is latitude, z is height, P is pressure, and a is the mean radius of the earth (6.37×10^6 m). (The hydrostatic assumption is used in expressing the right hand side in pressure, rather than Cartesian, coordinates).

The weighting in the longitudinal direction is simply given by the constant spacing between grid points in the λ direction, namely $\Delta\lambda = (2.5/360)(2\pi) = 4.363 \times 10^{-2}$ radians for 2.5° spacing. In the latitudinal direction, the simplest approximation is to replace the ϕ-dependent portion by $\cos(\phi(J)) \, \Delta\phi$ where $\phi(J)$ is the latitude indexed by J, namely $(37-J) \, \Delta\phi$, and $\Delta\phi$ is equal to $\Delta\lambda$. However, a slightly improved result can be obtained by viewing $\cos\phi \, d\phi$ as $d(\sin\phi)$, and deriving a set of coefficients appropriate to a trapezoidal rule quadrature in the unequally spaced variable $\sin\phi$. Imposing the conditions that this quadrature be exact for integrands which are constant or linear functions of $\sin\phi$, the coefficients are readily found to be given by $\cos(\phi(J)) \sin(\Delta\phi)$ at points away from the poles; this is very nearly equal to the approximate value since $\sin(\Delta\phi)/\Delta\phi = 0.99984 \approx 1$. At the poles, the coefficient is $0.5 \, (1-\cos(\Delta\phi)) = 0.47589 \times 10^{-3}$, compared with zero from the simpler approximation above.

In the vertical, trapezoidal rule weights are also employed, appropriate to the unequally spaced pressure levels. At levels below the uppermost (i.e. K<12) and above the surface, the weight is proportional to 0.5(P(K-1) - P(K+1)). Note that P(K-1) may refer to a mandatory pressure value or to the surface pressure, denoted PS(I,J). (The surface pressure is obtainable from the FGGE data using software developed at AFGL.) At the uppermost level (K=13, P(13) = 50mb), it is assumed the field values remain constant up to zero pressure; thus the appropriate weight is easily shown to be proportional to 0.5(P(12) + P(13)). At the surface, a K value (denoted KS(I,J)) is assigned thusly: if P(K) is the first mandatory pressure value that is smaller than PS(I,J) KS(I,J) is given the value K-1. (This necessitates denoting the mandatory levels from 2 to 13, rather than 1 to 12, since zero subscripts are not generally permitted in Fortran.) At a surface grid point, the vertical weighting is proportional to 0.5(PS(I,J) - P(KS(I,J)+1)). All grid points with K values less than KS(I,J) lie below the earth's surface and are given zero weight. (These vertical weight factors must be computed anew for each new surface pressure distribution).

Once the horizontal and vertical weight factors are combined, a representative expression for the weight (away from upper and lower boundaries and the poles) is

$$W(I,J) = 3.939\times10^{10}\cos(\phi(J))(P(K-1) - (P(K+1)) \qquad 1<J<73,\ KS(I,J)<K<13 \tag{6}$$

where the numerical factor is chosen so that if pressures are given in mb, W will be in kg. At the upper boundary,

$$W(I,J) = 2.36334\times10^{12}\cos(\phi(J)) \qquad 1<J<73,\ K=13 \tag{7}$$

while at the surface,

$$W(I,J) = 3.939\times10^{10}\cos(\phi(J))(P(KS(I,J) - P(K+1)) \qquad 1<J<73, \quad K=KS(I,J) \tag{8}$$

both of these expressions valid except at the poles. The expressions appropriate to the poles are obtained by replacing the $\cos(\phi(J))$ term by the numerical factor 0.01091. And finally, recall that W(I,J) vanishes for K<KS(I,J) (i.e., P(K)>PS(I,J)).

Once the weight field has been computed, the diagnostic computation can proceed fairly smoothly. The geopotential and temperature fields are given at the surface as well as at the mandatory pressure levels, hence the gravitational potential and internal energies can be easily computed and summed in accordance with (2) and (3). For the kinetic energy the horizontal velocity components at the surface are needed but not provided by the FGGE data. However approximate values can be obtained from the known values at the two closest mandatory pressure and the surface pressure value on the assumption that the velocity components are linear functions of ln P. (This approximation is consistent with the model and is in fact employed in the pre-and post-processing procedure.) It is then easy to perform the kinetic energy computation (1).

To compute the enstrophy, we first compute the local vertical component of the vorticity at each non-polar grid point:

$$\zeta = \frac{1}{a^2}\left[\frac{1}{\cos\phi}\frac{\partial V}{\partial \lambda} - \frac{\partial U}{\partial \phi}\right] \tag{9}$$

the derivatives being evaluated on a constant pressure surface. We employ the standard centered difference approximations to the horizontal derivatives, and hence must use the four values V(I±1, J) and U(I,J±1) at the level of the desired grid point in our calculation. If this grid point is a surface point, or if the surface pressure at any of the surrounding four points is less than that at the point under consideration, the necessary values (or pseudovalues) of U and/or V are obtained from the nearest neighboring pair of known values under the linearity assumption (in ln P) discussed above. Lateral variations in surface pressure resulting from changes in surface elevation are large enough so that velocities at five different levels must be stored in core simultaneously to allow required extrapolations; hence the $\partial V/\partial\lambda$ and $\partial U/\partial d\phi$ contributions to ζ must be computed in sequential steps. (The polar vorticities are of course calculated differently, utilizing the tangential velocity components at the neighboring latitude grid points: U(I,2) and U(I,72).) Once the vorticity field has been calculated, the computation of enstrophy (4) is straightforward.

A final computation performed on the FGGE grid is the global available potential energy, where the approximate form derived by Lorenz[6] is employed:

$$\text{Available potential energy} \doteq \frac{1}{2}\int_0^{PS} \frac{\bar{T}}{\gamma_d - \bar{\gamma}} \overline{\left(\frac{T'}{\bar{T}}\right)^2}\, dP \qquad (10)$$

where $\bar{(\ )}$ refers to a global average over an isobaric surface and $(\ )'$ is the departure from that average value. The lapse rate, $-\partial T/\partial z$,

is denoted by γ; γ_d is the dry adiabatic lapse rate (9.77×10^{3}°K/M). In evaluating the discreet approximation to (10), we have included surface values with others at the same K-level to compute isobaric averages and departures from average. The lapse rate is approximated (for given I,J) by:

$$\gamma(K) \doteq - \frac{T(K+1) - T(K-1)}{[\Phi(K+1) - \Phi(K-1)]/g} \qquad (11)$$

where the values at K-1 may represent surface values; at the lower and upper boundaries, one sided derivative approximations are employed. The computation of layer averages consists of taking weighted sums of the grid point values and dividing by the sum of grid point weights for the given layer. The integrand in (10) is then evaluated for each layer, and these values are summed to give the global available potential energy.

IV. MODEL GRID DIAGNOSTICS .

The computation of diagnostics in the model grid system is simpler for two important reasons. Firstly, the upper and lower boundaries are surfaces of constant σ (which is defined as the ratio of pressure to surface pressure; this means the top of the atmosphere is at σ=0 and the bottom is at σ=1 for all λ and ϕ.) Hence all the model grid points lie within the atmosphere. Secondly, the grid size (48 longitudinal points x 40 Gaussian latitudes x 6 σ-layers, or 11,520 total points) is small enough so that entire three dimensional arrays can be processed using core storage, thus eliminating the complexity of level by level computation.

The weights assigned to the grid points are calculated as follows. In the longitudinal direction, the weight is simply proportional to the uniform 7.5° (0.13090 radian) spacing. In the latitudinal direction, a 40 point Gaussian quadrature scheme is employed in the variable $\sin\phi$ (which takes on values +1 to -1 pole to pole); this fixes the 40 Gaussian values of $\sin\phi$ and hence ϕ, and also assigns the appropriate Gaussian weight to each set of latitudinal points. Tables providing these Gaussian weights are readily available[7]; we denote these values by GW(J).

In the vertical, the mass in each layer is proportional to the pressure difference across the layer which in turn is given by the layer thickness $\Delta\sigma(K)$ multiplied by the surface pressure PS(I,J). Hence the weight field can be written

$$W(I,J,K) = 5.42X10^{13}PS(I,J)GW(J)\Delta\sigma(K) \qquad (12)$$

where if surface pressure is in mb, W will be in kg.

Values of the velocity components, vorticity, geopotential and temperature at the model grid points are available, either from the FGGE data set via the preprocessing procedure, or from appropriate summation of the forecast fields in spectral form. Thus once the weight field has been calculated, the desired diagnostics (1) to (4) are readily computed. We have not yet established an appropriate formula for available potential energy with variables expressed in σ coordinates; once that is done, its evaluation on the model coordinate grid should be straightforward.

V. RECOMMENDATIONS

As a result of this summer's effort, a framework has been established within which to compute important atmospheric properties - kinetic and

potential energies and enstrophy - for the input and forecast fields of the AFGL primitive equation spectral model. After these computer codes have been implemented and checked, the aforementioned computations should be carried out, along with analagous spectral computations for the input and forecast fields. These results should be compared with each other and with computations on the actually observed fields. This will help answer questions such as how does pre- and post-processing affect energetics and how good is the model in predicting kinetic energy, and thus pinpoint areas in need of improvement.

It is further recommended that the pre- and post-processing procedures be carefully analyzed to determine the magnitude of errors introduced at these stages. For one thing, the transference of data from the 73 FGGE latitude to the 40 Gaussian latitude is accomplished by low order (essentially linear) interpolation. The effect of this process on the accuracy of the subsequent high order Gaussian quadrature should be studied. Further it is believed possible that the interpolation in the vertical to produce sigma level values may cause even greater inaccuracies. A thorough error analysis of these procedures, and of the initialization process, would help explain the results of the various model evaluation methods.

REFERENCES

1. J. Sela, "Spectral modeling at the National Meteorological Center," Mon. Weather Rev., Vol. 108, pp. 1279-1292, 1980.

2. G.W. Brier and R.A. Allen, "Verification of weather forecasts," Compendium of Meteorology, pp. 841-848, Boston, American Meteorological Society, 1951.

3. S. Teweles, Jr. and H.B. Wobus, "Verification of prognostic charts," Bull. Amer. Met. Soc., Vol. 35, pp. 455-463, 1954.

4. N.A. Phillips, "A coordinate system having some special advantages for numerical forecasting," J. Meteor., Vol. 14, pp. 184-185, 1957.

5. W. Bourke, B. McAvaney, K. Puri, and R. Thurling, "Spectral methods for atmospheric modeling," Methods in Computational Physics, Vol. 17, B. Adler, ed., New York, Academic Press, 1977.

6. E.N. Lorenz, "Available potential energy and the general circulation," Tellus, Vol. 7, pp. 157-167, 1955.

7. Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, M. Abramowitz and I.A. Stegun, eds., National Bureau of Standards Applied Mathematics Series No. 55, Washington, U.S. Government Printing Office, 1964.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINNERING EDUCATION

FINAL REPORT

A DIFFERENTIAL GEOMETRIC APPROACH TO ELECTROMAGNETIC LENS DESIGN

Prepared by: Dr. Alexander P. Stone

Academic Rank: Professor

Department and University: Mathematics Department, University of New Mexico

Research Location: Air Force Weapons Laboratory, Applied Physics Division, Electromagnetics Branch

USAF Research Colleague: Dr. Carl E. Baum

Date: August 2, 1982

Contract No: F49620-82-C-0035

ACKNOWLEDGMENT

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to spend a productive and very interesting summer at the Air Force Weapons Laboratory, Kirtland AFB, Albuquerque, New Mexico.

The laboratory, and in particular the electromagnetics branch in the Applied Physics Division, provided a very hospitable and excellent environment in which to carry out the research described in this report.

The author is very much indebted to Dr. Carl E. Baum for suggesting this particular area of research and for his guidance and helpful suggestions throughout the course of these investigations.

A DIFFERENTIAL GEOMETRIC APPROACH TO ELECTROMAGNETIC LENS DESIGN

By

Alexander P. Stone

ABSTRACT

A lens design technique developed by Carl E. Baum for transitioning TEM waves, ideally with no reflection or distortion, between cylindrical and conical transmission lines is investigated. This method uses a differential geometric approach combined with Maxwell's equations and the constitutive parameters ϵ and μ in an orthogonal curvilinear coordinate system. Isotropic but inhomogeneous media are considered. It is shown that rotational coordinate systems obtained from complex analytic transformations in the plane may be utilized in the design, and that a class of solutions to the design problem exists. This class of solutions is based on a Riccati type of differential equation. Suggestions for follow up research in this area are made.

I. INTRODUCTION:

The differential geometric approach to transient lens design is one of several promising approaches to electromagnetic problems. In general one starts with Maxwell's equations together with boundary conditions and general theorems such as conservation of energy and reciprocity and looks for various mathematical concepts for representing the solution of an EM problem. For example, operator diagonalization and the use of the complex frequency plane have proved to be extremely useful in the analysis and synthesis of EM devices. Other promising approaches, which remain to be thoroughly investigated, but which could have important application to EMP simulators and energy transport in pulse power equipment include topological properties of scatterers and group theoretic properties as well as the differential geometric method considered in this report. For the Air Force, these practical applications should be of great importance.

This report concerns inhomogeneous TEM plane waves which propagate on ideal transmission lines with two or more independent perfectly conducting boundaries. These types of inhomogeneous media can be used to define lenses for transitioning TEM waves, without reflection or distortions, between conical and cylindrical transmission lines. While there are practical limitations (e.g., the properties of materials used to obtain the desired permittivity and permeability of the inhomogeneous medium) perfect characteristics are not really necessary. This differential geometric approach to lens design was initiated by C. E. Baum (reference [1]) and further aspects of this method appeared in references [2] and [3]. Specifically, the differential geometric scaling method

creates a class of equivalent electromagnetic problems each having a complicated geometry and medium from an electromagnetic problem having a simple (Cartesian) geometry and medium. Included in these papers were examples of lenses which provided a perfect matching section between conical and cylindrical coaxial wave guides. In this paper a general design procedure for such lenses is specified and it is shown that the class of solutions of a certain Riccati equation yield a suitable lens design. This class includes in particular the examples given in [2]. Whether or not this class is unique is an open question at this time.

II. OBJECTIVES:

The main objective of this project was to study the differential geometry scaling procedure in EM lens design problems and obtain further examples of EM lenses. The uniqueness question, of when all such examples of lenses have been obtained, was also addressed but not answered completely. In order to obtain uniqueness it appears to be necessary to add certain conditions to the class of functions under study. The specific objectives were:

(1) To determine a suitable class of functions on which to base the design of EM lenses matching conical to coaxial waveguide.

(2) To determine specific examples of lenses.

III. The SCALING METHOD:

We first consider a Cartesian coordinate system (x,y,z) and an orthogonal curvilinear coordinate system (u_1,u_2,u_3) with line element

$$(ds)^2 = h_1^2(du_1)^2 + h_2^2(du_2)^2 + h_3^2(du_3)^2 \tag{1}$$

with the scale factors h_i , $i = 1,2,3$, given by

$$h_i^2 = \left(\frac{\partial x}{\partial u_i}\right)^2 + \left(\frac{\partial y}{\partial u_i}\right)^2 + \left(\frac{\partial z}{\partial u_i}\right)^2 \tag{2}$$

If certain combinations of the h_i , assumed positive, are defined as

$$(\alpha_{ij}) = \begin{pmatrix} h_1 & 0 & 0 \\ 0 & h_2 & 0 \\ 0 & 0 & h_3 \end{pmatrix}$$

$$(\beta_{ij}) = \begin{pmatrix} h_2h_3 & 0 & 0 \\ 0 & h_1h_3 & 0 \\ 0 & 0 & h_1h_2 \end{pmatrix} \tag{3}$$

and

$$(\gamma_{ij}) = \begin{pmatrix} h_2h_3/h_1 & 0 & 0 \\ 0 & h_1h_3/h_2 & 0 \\ 0 & 0 & h_1h_2/h_3 \end{pmatrix}$$

then one can write out the usual expressions for $\nabla\phi$, $\nabla\times\vec{E}$, and $\nabla\cdot\vec{E}$ in the u_i coordinate system in terms of physical components of the vector $\vec{E}$. Another set of vectors and operators (formal vectors and formal operators) may be written out in terms of tensor components.

For example, if $\vec{E}$ has components E_1 , E_2 , E_3 referred to the u_i coordinates, then $\vec{E}'$, the formal vector, has components E_1' , E_2' , E_3' and we write

$$\vec{E}' = (\alpha_{ij}) \cdot \vec{E} \tag{4}$$

while

$$\nabla\Phi = (\alpha_{ij})^{-1} \cdot \nabla'\Phi' \text{ , where } \Phi = \Phi' \text{ ,} \tag{5}$$

and

$$\nabla\times\vec{E} = (\beta_{ij})^{-1} \cdot \nabla'\times E' \text{ .} \tag{6}$$

The result one obtains is that Maxwell's equations

$$\begin{aligned} \nabla\times\vec{E} &= -\frac{\partial\vec{B}}{\partial t} \\ \nabla\times\vec{H} &= \vec{J} + \frac{\partial\vec{D}}{\partial t} \\ \nabla\cdot\vec{D} &= \rho \\ \nabla\cdot\vec{B} &= 0 \end{aligned} \tag{7}$$

together with the constitutive relations

$$\vec{D} = (\epsilon_{ij}) \cdot \vec{E}$$

and (8)

$$\vec{B} = (\mu_{ij}) \cdot \vec{H}$$

and the equation of continuity

$$\nabla\cdot\vec{J} = -\frac{\partial\rho}{\partial t} \tag{9}$$

can be rewritten in the form

$$\begin{aligned}
\nabla'\times\vec{E}' &= -\frac{\partial\vec{B}'}{\partial t} \\
\nabla'\times\vec{H}' &= \vec{J}' + \frac{\partial\vec{D}'}{\partial t} \\
\nabla'\cdot\vec{D}' &= \rho' \\
\nabla'\cdot\vec{B}' &= 0 \\
\vec{D}' &= (\varepsilon'_{ij})\cdot\vec{E}' \\
\vec{B}' &= (\mu'_{ij})\cdot\vec{H}' \\
\nabla'\cdot\vec{J}' &= -\frac{\partial\rho'}{\partial t}
\end{aligned} \tag{10}$$

where

$$\begin{aligned}
(\varepsilon'_{ij}) &= (\gamma_{ij})\cdot(\varepsilon_{ij}) \\
(\mu'_{ij}) &= (\gamma_{ij})\cdot(\mu_{ij}) \\
\text{and} \\
(\alpha'_{ij}) &= (\gamma_{ij})\cdot(\alpha_{ij})
\end{aligned} \tag{11}$$

where it is assumed that (ε_{ij}) , (μ_{ij}) , (α_{ij}) are real constant matrices which are independent of frequency, though possibly functions of position. Note that the primed equations (10) are of the same form as equations (7), (8), and (9), and so if we think of the u_i as a Cartesian coordinate system, a known solution of Maxwell's equations referred to Cartesian coordinates can be taken and if primed quantities are substituted for unprimed quantities, solutions to (10) can be found.

The result turns out to be that we have a solution to Maxwell's equations for which (ε_{ij}) , (μ_{ij}) , and (α_{ij}) may be anisotropic and/or inhomogeneous. The basic idea is then to pick (ε'_{ij}) , (μ'_{ij}) , and (α'_{ij}) and boundary surfaces of convenient forms in the u_i coordinate system so that a solution can be obtained in terms of the formal quantities. If some particular relationship between the u_i coordinates and the x,y,z coordinates is chosen, then the parameters (ε_{ij}) , (μ_{ij}) and the geometry of the boundary surfaces are determined and the solution can be applied to the case under study.

In this report we are concerned with problems related to inhomogeneous isotropic media and thus the constitutive parameter matrices are diagonal matrices of the form

$$(\varepsilon_{ij}) = \varepsilon(\delta_{ij}) \ , \ (\mu_{ij}) = \mu(\delta_{ij}) \tag{12}$$

where ε and μ are scalar functions of position. The formal quantities then look like

$$(\varepsilon'_{ij}) = \varepsilon(\gamma_{ij}) \ , \ (\mu'_{ij}) = \mu(\gamma_{ij}) \ . \tag{13}$$

We restrict $\alpha = 0$, so that the conductivity matrix is the zero matrix. Note that an inhomogeneous TEM wave with subscripts 1 and 2 only has no interaction with ε'_{33} or μ'_{33} in which case each of the matrices (ε'_{ij}) and (μ'_{ij}) has constant and equal diagonal entries in the first two diagonal positions. Hence such TEM solutions may be used to define lenses to match waves onto cylindrical and/or conical transmission lines.

IV. THEORETICAL DESIGN:

The (u_1,u_2u_3) coordinate system with the property that $h_1 = h_2$ is constructed as follows by considering any complex analytic transformation in the plane, say

$$z = F(w) = f(u,v) + jg(u,v) = \frac{x}{a} + j\frac{y}{a} \ , \ j = \sqrt{-1} \ . \tag{14}$$

The (x,y,z) coordinates are then obtained in the form

$$\begin{aligned} x &= f(u,v)\cos\phi \\ y &= f(u,v)\sin\phi \\ z &= g(u,v) \end{aligned} \tag{15}$$

by performing a rotation about the y-axis. The line element which results is

$$(ds)^2 = h_u^2(du)^2 + f^2(d\phi)^2 + h_u^2(dv)^2 \ . \tag{16}$$

where

$$h_u = \sqrt{(\frac{\partial f}{\partial u})^2+(\frac{\partial g}{\partial u})^2} \ . \tag{17}$$

Note that any plane orthogonal transformation, say $x = \alpha(u,v)$, $y = \beta(u,v)$ would yield on rotation about the y-axis the line element

$$(ds)^2 = h_u^2(du)^2 + \alpha^2(d\phi)^2 + h_v^2(dv)^2 \ . \tag{18}$$

Since one then may take $h_u = h_y$ without loss of generality (see reference [7]) for such rotational coordinate systems, the condition that α and β are the real and imaginary parts of an analytic function

is consequently forced upon us. Note also that because of the requirements on the constitutive parameters we must have the scale factors h_1 and h_2 equal for the u_1, u_2, u_3 coordinates. One may then show that a necessary and sufficient condition for this to occur is that the coordinate surfaces for constant v are spheres or planes if and only if the condition, $h_u/f(u,v)$ is independent of v, holds. A necessary and sufficient condition for this to occur is then found to be

$$(\frac{\partial f}{\partial u})(\frac{\partial^2 g}{\partial u^2}) - (\frac{\partial g}{\partial u})(\frac{\partial^2 f}{\partial u^2}) = (\frac{\partial g}{\partial u})h_u^2/f(u,v) \quad . \tag{19}$$

The u_i coordinates are then defined by setting

$$u_1 = \alpha_0(\exp \int_{u_0}^{u} \frac{h_u(u',v)}{f(u',v)}\, du')\cos\phi$$

$$u_2 = \alpha_0(\exp \int_{u_0}^{u} \frac{h_u(u',v)}{f(u',v)}\, du')\sin\phi \tag{20}$$

$$u_3 = F(v)$$

where $F(v) = \int_{v_0}^{v} h_u(\bar{u},v')dv'$, and $\bar{u} = 0$ if h_u is decreasing (for fixed v) in the range of interest for the coordinates (u,ϕ,v), while $\bar{u} = u_0$ if h_u is increasing (for fixed v) in the range of interest. It is also assumed that $g(u,0) = 0$ since this will guarantee continuity of conductors across one of the boundaries of the lens, say at $v = 0$. The corresponding scale factors are

$$h_1 = h_2 = (f(u,v))/(\exp \int_{u_0}^{u} \frac{h_u(u',v)}{f(u',v)}\, du') \tag{21}$$

and

$$h_3 = \frac{h_u(u,v)}{h_u(\bar{u},v)} \quad .$$

The scale factor h_3 will satisfy the condition that $h_3 \leq 1$ in the range of interest, and the equation (19) will guarantee continuity of the conductors across the other lens boundary at $v = v_0$. Knowledge of h_3 then permits calculation of the constitutive parameters for the lens.

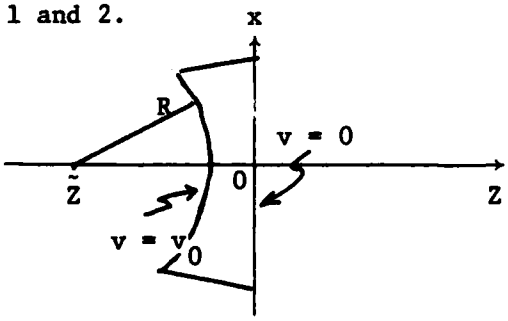
The radius of the sphere for the surface of constant v is

$$R = \left| f(u,v) h_u(u,v) / \frac{\partial g}{\partial u} \right| \tag{22}$$

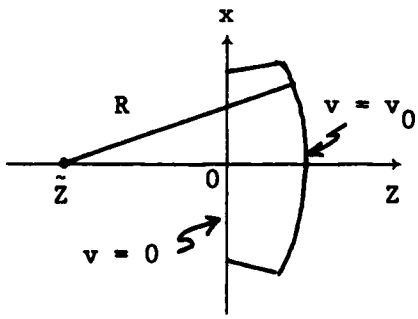
with its center located on the z-axis at

$$\tilde{z} = \left(f \frac{\partial f}{\partial u} + g \frac{\partial g}{\partial u} \right) / \frac{\partial g}{\partial u} \quad . \tag{23}$$

Both R and $\tilde{z}$ depend only on v and hence are constants for the spheres corresponding to constant v. The lens is classified as convergent if $\tilde{z}^2 > R^2$ and $\frac{\partial h_u}{\partial u} > 0$ in the range of interest, and divergent if $\tilde{z}^2 < R^2$ and $\frac{\partial h_u}{\partial u} < 0$ in the range of interest. See Figures 1 and 2.



Convergent Lens
Figure 1



Divergent Lens
Figure 2

V. A CLASS OF SOLUTIONS:

In order to obtain a class of solutions for a lens design, the condition given in equation (19) is one that must be satisfied as well as the condition that $g(u,0) = 0$. For a convergent lens we must also satisfy $\tilde{z}^2 > R^2$ and $\frac{\partial h_u}{\partial u} > 0$. It can be shown directly that all solutions of the Riccati differential equation

$$\frac{dz}{dw} = \frac{1}{2}(1+z^2) \tag{24}$$

satisfy these conditions. The family of solutions is given by

$$z = \tan(\frac{w+k}{2}) \ , \tag{25}$$

where k is a constant.

Similarly, in the case of a divergent lens the solutions

$$z = \tanh(\frac{w+k}{2}) \tag{26}$$

of

$$\frac{dz}{dw} = \frac{1}{2}(1-z^2) \tag{27}$$

will generate a class of divergent lenses. The case $k = 0$ in equations (25) and (26) corresponds directly to the examples which appear in references [1] and [2]. Other examples may then be generated by varying k .

Thus if we set $p = z/a$, $q = w/b$, the differential equations (24) and (27) may be replaced by the single Riccati equation

$$\frac{dp}{dq} = \frac{1}{2}(1+p^2) \quad . \tag{28}$$

and this equation, along with its solutions, will generate a class of lenses, either divergent or convergent. In particular, if one sets $p = jz$, $q = jw$ in equation (28), then both examples of [1] are obtained.

VI. RECOMMENDATIONS:

The question that arises naturally from section V is whether or not all solutions to the design problem for a lens matching a conical to a cylindrical transmission line have been obtained. The main condition to be satisfied by an analytic function $z = F(w) = f(u,v) + jg(u,v)$ is given by equation (19). One may impose some additional requirements on $F(w)$ and find that the differential equation (28) is satisfied, but the question is are these natural? Thus uniqueness is a central question.

Other recommendations for follow-on research would be (a) to consider further examples in which the medium is allowed to be anisotropic as well as inhomogeneous, (b) to consider the possibility of removing the orthogonality requirement for our coordinate systems, and (c) determination of a symmetry group for which the lens problem considered here is a subgroup.

Finally, in reference [1] several examples of two-dimensional lens are given. This simpler case in which lenses appropriate for launching TEM waves between wide perfectly conducting parallel sheets should not be overlooked as a source of more examples for lenses.

In summary, the importance of differential geometry in transient

EM theory seems clear from the results which have been obtained by C. E. Baum in his many investigations and also from the work which is reported here. In addition, certain techniques from differential geometry, such as the use of differential forms in EM theory, have been receiving attention in recent work. These are positive indications that methods from differential geometry if properly exploited will be one of possibly several fruitful ways modern mathematics will contribute to the development of transient and broad-band electromagnetics.

REFERENCES

1. C. E. Baum, "A scaling technique for the design of idealized electromagnetic lenses", Sensor and Simulation Note #64, August 1968.

2. T.C. Mo, C.H. Papas, and C.E. Baum, "Differential geometry scaling method for electromagnetic field and its applications to coaxial waveguide junctions", Sensor and Simulation Note #169, March 1973. (Also as a paper by the same authors under the title of "General scaling method for electromagnetic field with application to a matching problem", J. Math Physics, Vol. 14, pp. 479-483, April 1973.)

3. C.E. Baum, "Emerging Technology for Transient and Broad-Band Analysis and Synthesis of Antennas and Scatterers", Proceedings of the IEEE, vol. 64, #11, pp. 1598-1616, November 1976.

4. P. Moon and D.E. Spencer, Field Theory for Engineers, D. Van Nostrand, Princeton, New Jersey (1961).

5. J.J. Stoker, Differential Geometry, Wiley-Interscience, New York (1969).

6. R.V. Churchill, Complex Variables and Applications, 2nd edition, McGraw-Hill, New York (1960).

7. P. Moon and D.E. Spencer, "Cylindrical and rotational coordinate systems", J. Franklin Institute, vol. 252, pp. 327-344, 1951.

8. I.S. Sokolnikoff, "Tensor Analysis Theory and Applications to Geometry and Mechanics of Continua, 2nd edition, John Wiley and Sons, New York (1964).

1982 USAF - SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

THE USE OF EXPERIMENTAL INPUT IN TRANSONIC AERODYNAMICS

Prepared by: Dr. Alfred G. Striz

Academic Rank: Assistant Professor

Department and University: School of Aerospace, Mechanical and Nuclear Engineering
University of Oklahoma

Research Location: Air Force Armament Laboratory,
Munitions Division,
Aircraft Compatibility Branch

USAF Research Colleagues: Robert W. Bunton, Dave M. Belk

Date: September 8, 1982

Contract No.: F49620-82-C-0035

THE USE OF EXPERIMENTAL INPUT IN TRANSONIC AERODYNAMICS

by

Alfred G. Striz

ABSTRACT

A brief description of various two-dimensional and three-dimensional steady and unsteady transonic codes is given which are evaluated for possible implementation on the Eglin AFB computer system. A modifiction is presented to the two-dimensional small-disturbance transonic code STRANS2 to accept experimental pressure distributions as input. This will generate a semi-empirical output to be used as input to the harmonic-analysis unsteady transonic code UTRANS2. The aerodynamic coefficients thus obtained can be used in a strip-theory-type analysis of three-dimensional wings. Also, the unsteady linear transonic Kernel function program ANKF is being evaluated by comparison to experimental results obtained at NLR. This code utilizes experimental steady Mach number distributions and mode shapes as input. For a subsonic case, the code shows good comparison. Only limited comparison is obtained for a transonic case. Finally, a short description is given of the contents of a seminar on Fundamentals of Aeroelasticity which was conducted at AFATL/DLJCS.

ACKNOWLEDGEMENT

The author would like to express his appreciation to the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to spend an interesting and stimulating summer at the Air Force Armament Laboratory, Eglin AFB, FL. He would like to thank the laboratory, especially Dr. Don Daniel, and the Aircraft Compatibility Branch under John Macon for the excellent working conditions and the friendly hospitality provided.

Finally, he would like to thank the Flutter Group for making this summer rewarding and enjoyful, espcecially Bob Bunton and Dave Belk for suggesting this area of research and for their help and support, and Jim Robinson for many interesting discussions.

I. INTRODUCTION

Most airplanes face an increased risk of flutter in the transonic flight regime. Thus, theoretical transonic aeroelasticity is vital for the flutter evaluation of supersonic airplanes, especially advanced fighter aircraft. However, the computation of steady and unsteady transonic flows over airfoils and wings with and without tip and underwing stores poses distinct problems. Despite recent advances in computer codes, severe restrictions still apply both with respect to the computer environment and the accuracy of the computed results.

Most of these problems are caused by the non-linearity of the flow which is partly subsonic and partly supersonic. The finite difference and finite element schemes used to solve these equations are iterative in nature and, combined with a large computational grid, result in tremendous demands on computer time and memory. Also the aforementioned difficulties influence the accuracy of the solutions.

II. OBJECTIVES OF THE RESEARCH EFFORT

Thus, one of the objectives of this summer research was the evaluation of two-dimensional and three-dimensional transonic codes for possible implementation on the CDC computers at Eglin AFB. A short description of this effort is given in Section III of this report.

In the course of this evaluation it became obvious that present three-dimensional codes seem to be too large in both time and memory requirements to be reasonably used for extensive computation of aerodynamic coefficients for transonic flutter analysis of complex aircraft. Thus, alternative and less complex ways have to be found to achieve three-dimensional transonic flow solutions.

One such attempt could be the use of two-dimensional transonic flow codes such as STRANS2/UTRANS2[1] in strip-theory for aeroelastic applications. However, as the spanwise flow is not accounted for in this case, a purely theoretical solution would most likely misrepresent the true conditions. To more closely model the real three-dimensional conditions it was suggested to use experimental input from three-dimensional tests to the steady part of the two-dimensional small perturbation transonic code in order to improve the unsteady solutions of the same program. A description of this attempt is given in Section IV. Also, the possibility of using a three-dimensional linear but transonic Kernel function unsteady code like ANKF[2,3] was investigated. This code, too, depends on experimentally obtained steady flow data and structural or aeroelastic mode shapes as input. It is, however, far less complex and needs only fractions of the computer time and space requirements of three-dimensional non-linear transonic codes. Section V contains the results of this study.

Finally, Section VI gives a short overview of the subject areas covered by the author in a seminar on **Fundamentals of Aeroelasticity** taught during the summer period.

III. EVALUATION OF 2-D AND 3-D TRANSONIC CODES

The number of new or modified transonic steady and unsteady flow codes has been growing steadily in the last decade. A review of the state-of-the-art in transonic flow can be found in Reference.[4]

Three-Dimensional Transonic Codes

Some of the most promising 3-D transonic steady and unsteady computer programs were evaluated for possible implementation on the Eglin AFB CDC-computing system. **LTRAN3** is a three-dimensional unsteady transonic code that solves the small-disturbance potential equation in the low-frequency range by the time-integration method. The program was developed by Borland, Rizzetta and Yoshihara[5] out of the equivalent two-dimensional code LTRAN2 by Ballhaus and Goorjian[6]. A comparison with experimental data from NASA for an unswept rectangular flexible wing with circular-arc airfoil sections was performed by Guruswamy and Goorjian[7] with good comparison for a subsonic case. In a strongly transonic case, the results compared well except at the wing root, where strong viscous effects were not accounted for. Despite its large computer memory and time requirements, this program could be implemented on the Cyber 176 at Eglin AFB. However, it is not supported by NASA at this time due to the development by Borland and Rizzetta of the modified code **XTRAN3S**, which includes high-frequency terms in both the governing differential equation and the boundary conditions[8]. Also, a static and dynamic aeroelastic capability has been added in this program. However, it has not yet been compared to experimental results. As this code is exclusively supported on the CRAY 1S computer due to its large size and long CPU-times, the only possible use to AFATL would be by telephone hook-up through a dial-up modem.

A somewhat less complex solution to the transonic flow problem is given by Traci, Albano, and Farr[9] in their three-dimensional steady and

unsteady small-disturbance codes, **TDSTRN and TDUTRN.** Here, the unsteady program TDUTRN is based on harmonic analysis method where the unsteady solution is linearized with respect to time, limiting it to small oscillations. Eastep and Olsen[10] utilized these codes in a flutter analysis of a rectangular wing with conventional airfoil sections. Again, no comparison to experimental results has yet been shown. This set of codes is presently being verified for release at AFWAL and could with relative ease be implemented on the Cyber 176 for use by AFATL.

One of the most advanced steady transonic flow codes was developed by Boppe[11]. The finite difference scheme used is characterized by a multiple embedded mesh system. The code can treat complex realistic aircraft shapes by use of an extended small-disturbance flow equation. Also, correlation with some experimental data has been shown. The transonic steady results from this program could possibly be used as input to some linear unsteady program for computation of aerodynamic coefficients. Presently, the computer code is being evaluated at AFATL/DLJCA.

A different approach to the solution of the transonic flow problem is taken by Tseng and Morino[12] who presented an extension of the Green's function method to non-linear unsteady transonic flows. This modification of the integral equation method is expected to perform one order of magnitude faster than finite difference schemes, also allowing for complex configurations and high reduced frequencies. The transonic extension together with an aeroelastic capability is presently being written into a production code, **SUSAN**, which is based on a similar but not

transonic code SOUSSA-P1.1[13]. Implementation at AFATL/DLJCS is expected.

Two-Dimensional Transonic Codes

The two-dimensional transonic steady and unsteady flow codes **LTRAN2-HI**[14] and **STRANS2/UTRANS2**[1] were implemented on the CDC-computers at Eglin AFB and test cases were run for verification. LTRAN2-HI is a version of LTRAN2 which was modified by Hessenius and Goorjian to include high-frequency time-dependent terms in the calculation of the pressure coefficient as well as in the wake and downstream boundary conditions. STRANS2 and UTRANS2 are the two-dimensional predecessors to the codes TDSTRN and TDUTRN mentioned earlier. Both LTRAN2 and STRANS2/UTRANS2 were used in extensive flutter analyses of various airfoils by Yang, Guruswamy, and Striz [15-17].

Despite their inherent limitations (no high-frequency term in the governing equation of LTRAN2-HI, and harmonic analysis method with convergence problems for higher reduced frequencies in UTRANS2), these two programs could be used in a strip-theory-type analysis for known mode shapes to be compared to three-dimensional theoretical or experimental results.

Other two-dimensional transonic steady and unsteady flow programs of interest include **LTRAN2-NLR**[18], **HYTRAN2**[19], **EXTRAN2**[20], and a code by Isogai[21].

In LTRAN2-NLR, Houwink and van der Vooren added the high-frequency terms in the boundary conditions to LTRAN2. Edwards, Bennett, Whitlow, and Seidel further modified LTRAN2-NLR by using a monotone differencing

scheme to eliminate expansion shocks near the airfoil leading edge. EXTRAN2 by Rizzetta and Chin is a different modification of LTRAN2-NLR which incorporates the high-frequency term in the governing equation. Finally, Isogai developed a transonic small-perturbation code which utilizes a time-marching, semi-implicit and implicit two-sweep procedure as finite difference scheme.

THREE-DIMENSIONAL LINEAR CODES

Another alternative to the use of complex transonic non-linear codes might be the use of a linear Kernel function program like ANKF by Cunningham[2,3]. In the transonic case, this unsteady program uses steady experimentally or theoretically obtained Mach number distributions and known mode shapes as input. The shock is treated as doublet-line on the wing surface. This program is being evaluated in comparison to experimental results in Section V.

IV. MODIFICATION OF STRANS2 FOR USE OF EXPERIMENTAL INPUT

The computer programs STRANS2 and UTRANS2 developed by Traci, et. al. are based on finite difference schemes to solve the steady and unsteady small-disturbance transonic aerodynamic equations, respectively. The steady state solution is obtained by a mixed differencing relaxation procedure. The unsteady solution is obtained by using the same procedure following a harmonic analysis method. The unsteady equation is linearized with respect to time.

Because UTRANS2 uses the steady potential at each grid point as input to calculate the unsteady potentials, accuracy of the steady potential solution obtained from STRANS2 greatly affects the unsteady

solution. Due to the inherent limitations in the nonlinear steady program, i.e., the numerical errors induced by the finite difference scheme (particularly in the treatment of shocks), the absence of viscosity terms in the flow equation, possible inaccuracy in the airfoil shape functions, etc., the STRANS2 results are not often in good agreement with experimental data. The lack of agreement is especially pronounced for higher transonic Mach numbers and larger angles of attack, when stronger shocks are present. To improve such lack of agreement, it was suggested to modify STRANS2 so that is can use experimental steady pressure data along the airfoil surfaces as input to generate potentials at all grid points. STRANS2 now uses airfoil shape functions as input. With the continuing activities in transonic wind tunnel and flight tests, experimental pressure data become increasingly available for transonic Mach numbers less than one.

Numerical Scheme Used in STRANS2

The following boundary value problem is solved in STRANS2 to compute the steady velocity potential ϕ^0 at each grid point.

The transonic steady flow equation is

$$(K - \phi_x^0)\,\phi_{xx}^0 + \phi_{yy}^0 = 0$$

where K is the transonic similarity parameter.

The boundary conditions on the airfoil are

$$\phi_y^0 = f_{u,\ell}^{0'}(x) \quad \text{on } y = \pm 0,\ 0 \le x \le 1 \tag{2}$$

where $f_{u,\ell}^{0'}(x)$ represent the slopes of the upper and lower surfaces on the airfoil.

A wake condition and a farfield boundary condition complete the problem.

In the numerical scheme, Eq. (2) is differentiated to give ϕ^0_{yy}, which is substituted into the steady flow equation (1) for calculation of potentials on the airfoil surfaces.

Modifications

The necessary modifications consist of two parts: (i) to compute the steady velocity potentials along the airfoil surfaces from experimental pressure data; and (ii) to replace the portions of STRANS2 that treat the airfoil surface boundary condition by portions that use the potentials on the airfoil surfaces.

(i) The steady pressure C_p^o and the velocity potential ϕ^0 are related as

$$C_p^0 = -2 \frac{\delta^{2/3}}{\{(1+\gamma) M_\infty^2\}^{1/3}} \quad \phi_x^0 = -S\phi_x^0 \tag{3}$$

Based on the experimental steady pressure data a curve fitting scheme is used to derive functions describing the pressure distribution along the upper and lower surface of the airfoil. In the presence of a shock, an upstream and a downstream function have to be derived. These functions are in the symbolic form

$$g^0_{u,\ell}(x) = \phi_x^0 = -C_p^0/S \tag{4}$$

Thus the function for the steady potential is obtained by integrating Eq. (4),

$$\phi^0_{u,\ell}(x) - \phi^0_{u,\ell}(0) = \int_0^x \frac{\partial \phi^0_{u,\ell}(\xi)}{\partial \xi} d\xi = \int_0^x g^0_{u,\ell}(\xi)\, d\xi \tag{5}$$

where $\phi^0_{u,\ell}(0)$ are the two constants of integration. The condition at

the leading edge requires that the potential on the upper and lower surface be equal. Thus, Eq. (5) actually contains one constant, $\phi^0(0)$.

(ii) The slope data $f^{0'}_{u,\ell}(x)$ along the airfoil boundaries can be related to velocity potentials as

$$f^{0'}_u(x) = \frac{\phi^0_{i,j} - \phi^0_{i,j-1}}{\Delta y_{j-1}}; \qquad f^{0'}_\ell(x) = \frac{\phi^0_{i,j+1} - \phi^0_{i,j}}{\Delta y_j} \tag{6}$$

The potentials $\phi^0_{i,j-1}$ on the upper surface and $\phi^0_{i,j+1}$ on the lower surface are known quantities from Eq. (5). However, $\phi^0_{i,j}$ at a grid point above the upper surface and $\phi^0_{i,j}$ at a grid point below the lower surface are unknown. Since the $\phi^0_{i,j}$'s are already unknown in the finite difference equations, no additional unknowns are introduced.

After obtaining a converged solution of the potential ϕ^0 at every grid point, it can be put into UTRANS2 for computation of the unsteady velocity potentials.

Revision of STRANS2

The following example was used in implementing the changes to STRANS2.

The steady pressure coefficient distribution for a NACA 64A010 airfoil at M = 0.802 and α = 0.0° was taken from experimental results obtained by Davis and Malcolm in Reference 22 (Fig. 13, D.I. = 12). The distribution is shown in Figure 1 together with a steady pressure curve obtained by STRANS2 for the same airfoil at M = 0.80 and α= 0.0°.

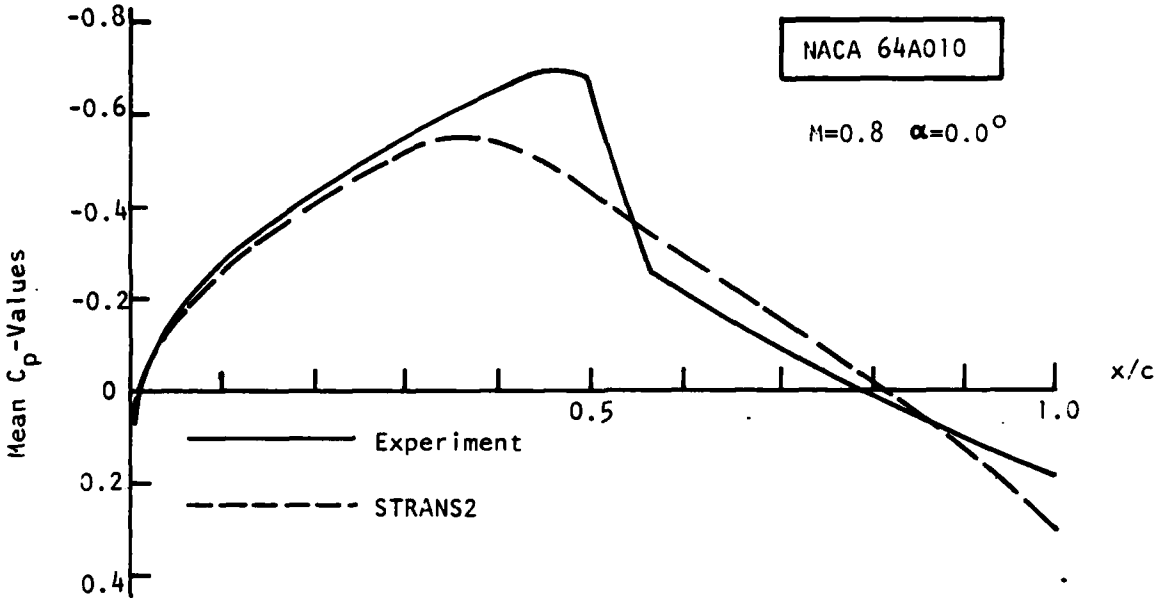


Figure 1. Steady Pressure Distribution for NACA 64A010 Airfoil

The C_p^o-values at each station x on the airfoil surface were divided by the parameter (-S) to give the function values $g^0_{u,\ell}(x)$. Because of symmetry of the airfoil, the function values on the upper and lower surface are identical in value for $\alpha = 0.0°$. The functions $g^0_{u,\ell}(x)$ were then divided into two parts: one from the leading edge to the approximate shock position; the other from the shock to the trailing edge. A least-squares-fit program was used to obtain the function values at each station x on the airfoil.

The values for the steady velocity potential at these stations were computed by an integration program which utilizes an adaptive extrapolation scheme to solve Eq. (5), where the reference potential $\phi^0(0)$ was evaluated from the original version of STRANS2 computing the same flow for the same airfoil.

The changes to STRANS2 were thus limited to the main program, one subroutine and two functions. These changes were incorporated into the code, which was then called STREX2.

Results

In the course of this investigation, convergence could not be achieved on STREX2. It seems that the steep slopes of the experimental input might be responsible for the divergence. A variation in relaxation parameters was run and parameters found which slowed divergence but did not lead to convergence. It is suggested to change the derivatives in the slope-potential relationships (Eqs.6) to a more complex finite difference scheme. This work is presently being continued at the University of Oklahoma.

V. EXPERIMENTAL INPUT TO A LINEAR UNSTEADY TRANSONIC 3-D KERNEL FUNCTION CODE

An alternate way of performing transonic unsteady flow computations while improving on the computer memory and time requirements is the use of linear codes like ANKF by Cunningham[2,3].

Description of Code

This program represents a steady and oscillatory Kernel function method for subsonic, transonic, and supersonic flow. The theory is based on linear lifting surface theory and makes the program applicable to multiple interfering coplanar, non-coplanar, or intersecting planar surfaces. For transonic cases with imbedded shocks, the non-uniform unsteady flow is treated in the form of a small linear perturbation

about non-linear flow fields. These flow fields are input from either experiment or other, finite-difference-type, solutions as steady local Mach number distributions on the surfaces together with structural or aeroelastic mode shapes. In the transonic analysis, results have to be obtained separately for the upper and lower surfaces as these solutions are assumed to be decoupled. The trailing edge and wing tip pressure coefficients are set to zero for small perturbances, whereas the leading edge pressures can be arbitrary. The shock is simulated by a line-doublet between two interfering surfaces, an all supersonic or subsonic to supersonic upstream surface and a subsonic downstream surface. Generalized forces and pressure distributions can be calculated in unsteady flow.

This code was evaluated to determine the influence of the experimental input data on subsonic and transonic flow results by comparison with experimental unsteady results.

Experimental Data

All experimental steady input and unsteady comparison data were taken from reports by NLR, the National Lucht-en Ruimtevaartlaboratorium of the Netherlands[23]. In these reports, the authors present transonic wind-tunnel tests on an F-5 wing (Fig 4 of Ref. 23, Part I) with and without an external store oscillating in a pitch mode at various Mach numbers from M = 0.6 to 1.35 and frequencies up to 40 Hz. Results are presented as pressure distributions on upper and lower surfaces for the clean wing, the wing with tip-mounted store, and with underwing-mounted store. The mode shapes are also presented as displacement functions

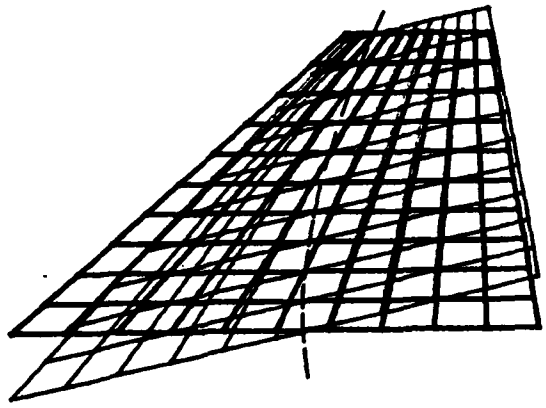


Figure 2. Pitch Mode for F-5 Wing at M=0.95 and 20Hz

assuming no deformation in chordwise direction and a parabolic deformation in spanwise direction. An example for a pitchmode at M = 0.95 and 20 Hz is given in Fig. 2. The nodal line is shown as dashed line.

The test runs picked for this comparison were those at M = 0.6, 0.8, and 0.95 at 20 and 40 Hz, respectively. The non-uniform Mach number distributions for these cases were obtained and interpolated to the respective input stations: 6 x 6 uniformly distributed stations for the subsonic cases at M = 0.6 and 0.8, and 5 x 6 and 3 x 6 uniformly distributed stations on the respective upstream and downstream surfaces in the transonic case of M = 0.95. At the trailing edge and the wing tip the freestream Mach number was assumed. The mode shapes for the mentioned test runs were obtained and plotted.

Other input parameters to the code that can be varied include surface type (chosen: type 3, a single surface detached from the fuselage center line), spanwise loading (chosen: type 3, detached from the fuselage center line, decreasing as $\sqrt{1-\bar{\eta}}$, where $\bar{\eta}$ is the non-dimensional

distance along the span away from the wing root toward the tip), and chordwise loadings. The influence of the latter was evaluated.

Results

The following conclusions can be drawn with respect to the resulting pressure jump distributions (ΔC_p).

For all cases run (M = 0.6, 0.8, and 0.95 at frequencies of 20 Hz and 40 Hz), the comparison between the results from ANKF and the experimental results improved from the wing root toward the wing tip. At the innermost station (18% span), the Kernel function method tended to underpredict the experimental results considerably. This may be due to the fact that it does not account for viscous and fuselage/wall interference effects. Toward the wing tip, this underprediction decreased to rather good comparison for the subsonic cases. This was true for both, real and imaginary, parts of the pressure coefficients. With increase in Mach number the comparison seemed to become less favorable. At the transonic Mach number of M = 0.95, the Kernel function results are seen to overpredict on the upstream portion of the wing while they underpredict the pressures on the downstream portion (Fig. 3). Also, the pressure representation is somewhat unclear due to two distinct lift forces for the shock movement of the upper and lower surface, respectively, versus a single shock for the experimental data. The increase in frequency from 20 Hz to 40 Hz did not seem to effect comparison noticeably.

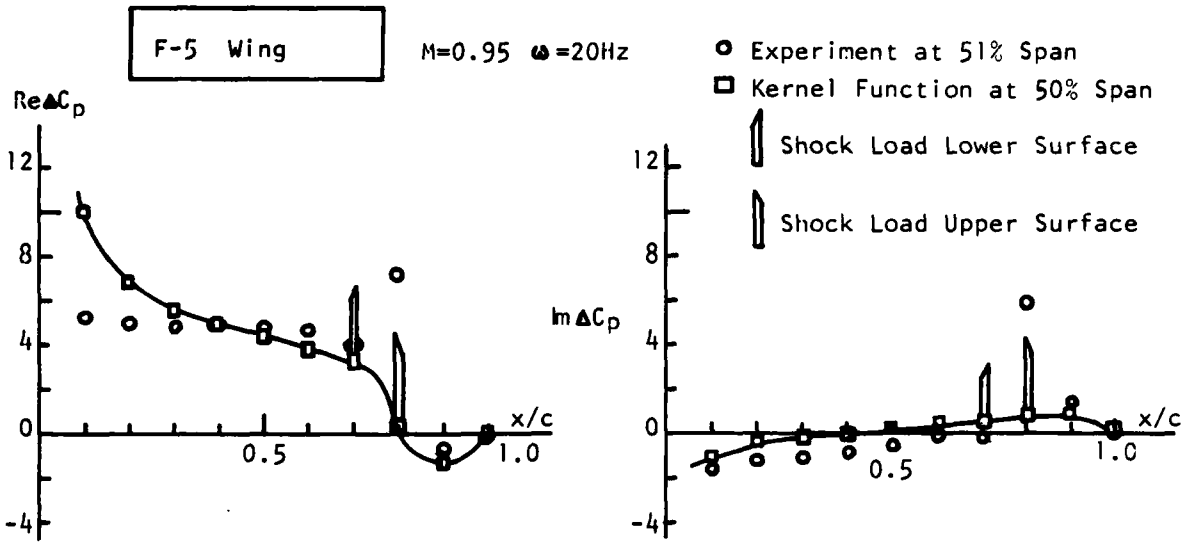


Figure 3. Unsteady Pressure Distribution on F-5 Wing

Use of the chordwise loadings available (Fig. 4) showed the following: results for loading 3 are similar to those for loading 1 with higher pressures toward the leading edge and lower pressures toward the trailing edge; results for loading 4 are similar to those for loading 2 with nearly the same pressures toward the front portion of the wing and higher values toward the downstream portion. The two loadings 2 and 4 show higher upstream pressures and lower downstream pressures than loadings 1 and 3.

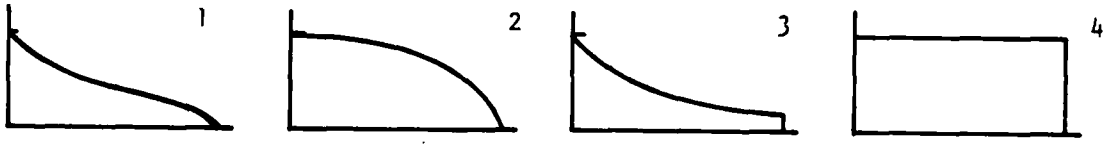


Figure 4. Chordwise Loading Types for ANKF-Code

In order to determine the influence of the experimentally obtained input data, a comparison was made between cases with uniform freestream

Mach number over the whole surface and cases with the experimental local Mach number distribution. The results showed that the use of the experimental Mach numbers improved the results by increasing pressures on the upstream part of the airfoil while decreasing those on the downstream portion. Also, a case was run with a very small amplitude of oscillation. The resulting pressures were very small but showed essentially the same distribution as for large amplitudes. Thus, the influence of the mode shape is rather distinct. The influence of the leading edge Mach number on the pressure distribution was determined from the comparison of runs with freestream Mach number at the leading edge versus somewhat higher experimentally determined Mach numbers. In all cases, the experimental data tended to increase the pressures on the wing favorably.

This investigation should be extended to more Mach numbers and more reduced frequencies for better evaluation of the Kernel function code. Also, cases with added tip-store should be treated and compared to experimental results from NLR. The tipstore should be modelled as a flat plate and also as a vertical endplate to evaluate the quality of the respective tip-store model and to investigate the influence of the interfering surfaces on each other. This will be explained in more detail in Section VII.

VI. SEMINAR ON FUNDAMENTALS OF AEROELASTICITY

During the course of the summer a seminar was conducted on **Fundamentals in Aeroelasticity** which originated from a graduate course in aeroelasticity taught by the author at the University of Oklahoma.

The following subjects were treated (e.g. Ref. 24, 25).

-Steady Aeroelasticity

After the development of structural theorems (covering mainly bending and torsion equations) and of steady and unsteady aerodynamics (including wing theories and the wake influence on lift and moments), solution procedures were given for the aeroelastic equation. Divergence and rolling effectiveness were presented in detail.

-Transient Aeroelasticity

The gust problem (Wagner's Problem) was explained.

-Unsteady Aeroelasticity

Following the derivation of the equations of motion, the aerodynamic forces and moments were obtained and flutter equations were presented for both two- and three-dimensional cases. Various flutter solution procedures (assumed mode or Rayleigh-Ritz method, U-g, p and p-k methods) were elaborated on. With regard to transonic aeroelasticity, the influences of the shock and the shock movement were discused. Also, panel flutter and related solution methods were treated.

VII. RECOMMENDATIONS

At this point, the following recommendations with regard to the performed work can be made:

1. A two-dimensional strip-theory-type investigation should be performed utilizing one of the more advanced 2-D codes like LTRAN2-HI to compute steady and unsteady pressure distributions and

aerodynamic coefficients at various spanwise stations on a wing. These results should be compared to existing ones from a three-dimensional code like LTRAN3 and also to experimental results for a three-dimensional wing.

2. The modification of STRANS2 to accept experimental input should be continued to result in the semi-empirical code STREX2. The output from both STRANS2 and STREX2 should then be used as input to UTRANS2 and the results for both cases should be compared with each other and with experimental two-dimensional results. If these comparisons are favorable, STREX2/UTRANS2 should be utilized for an analysis like that described in the previous paragraph. The modification of STRANS2 will be continued by the author at the University of Oklahoma.

3. The steady and unsteady Green's function transonic code SUSAN seems promising. Its algorithm should be verified and the code evaluated by comparison to experimental data. Also, the aeroelastic capability of SUSAN should be compared to a large production code like the aeroelastic subprogram in NASTRAN/MSC.

4. More testcases should be run on the transonic unsteady Kernel function code ANKF possibly for various other wings, more Mach numbers, more reduced frequencies, and various mode shapes to gather more data for comparison with other codes (e.g. panel method or Doublet-Lattice method) and experimental results. Furthermore, the influences of the mode shape, the experimental Mach number distribution, and the choice of surfaces in the transonic case on the unsteady

results should be investigated in more detail. Additional steady transonic input to ANKF could be obtained by a 3-D code such as Boppe's or by other transonic steady codes. Also, the addition of a tip-store to the clean wing should be analyzed and the results compared to experimental data.

5. During the course of this research it became obvious that in most cases of practical application of the Kernel function or Doublet Lattice method to the aerodynamics of wings with attached tip-stores, the stores are modelled as flat plates for simplicity. This does not seem to account for the inhibition of spanwise flow at the wind tips due to the three-dimensional geometry of the attached body. Thus, various combinations of a flat plate with a vertical endplate for spanwise flow control should be analyzed. The results should be compared to experimental data for wings with tip-stores[22] and to results for a full three-dimensional tip-store model. In this way, an optimal model could be established with regard to simplicity and accuracy.

REFERENCES

1. Traci, R.M., Albano, E.D., and Farr, J.L., **Perturbation Method for Transonic Flows About Oscillating Airfoils**, AIAA Journal, Vol. 14, Sept. 1976, pp. 1258-1265.

2. Cunningham, A.M. Jr., **A Steady and Oscillatory Kernel Function Method for Interfering Surfaces in Subsonic, Transonic and Supersonic Flow**, NASA CR-144895, Sept. 1976.

3. Cunningham, A.M. Jr.,and Anderson, J.D., **Supplementary Notes on Appendix H of NASA CR-144895**, NASA/Langley Research Center, Purchase Order No. L 72474A

4. Ballhaus, W.F., Deiwert, G.S., Goorjian, P.M., Holst, T.L., and Kutler, P., **Advances and Opportunities in Transonic Flow Computations**, Numerical and Physical Aspects of Aerodynamic Flows, Springer, 1981.

5. Borland, C.J., Rizzetta, D.P., and Yoshihara, H., **Numerical Solution of Three Dimensional Unsteady Transonic Flow Over Swept Wings**, AIAA Paper 80-1369, Snowmass, Colo., July 1980.

6. Ballhaus, W.F. and Goorjian, P.M., **Implicit Finite Difference Computations of Unsteady Transonic Flows About Airfoils**, AIAA Journal, Vol. 15, Dec. 1977, pp. 1728-1735.

7. Guruswamy, P. and Goorjian, P.M., **Comparisons Between Computations and Experimental Data in Unsteady Three-Dimensional Transonic Aerodynamics, Including Aeroelastic Applictions**, AIAA Paper 82-0690, New Orleans, Louis., May 1980.

8. Borland, C.J. and Rizzetta, D.P., **Nonlinear Transonic Flutter Analysis**, AIAA Paper 81-0608-CP, Atlanta, GA, April 1981.

9. Traci, R.M., Albano, E.D., and Farr, J.L., **Small Disturbance Transonic Flows About Oscillating Airfoils and Planar Wings**, AFFDL-TR-75-100, Air Force Flight Dynamics Lab., June 1975.

10. Eastep, F.E. and Olsen, J.J., **Transonic Flutter Analysis of a Rectangular Wing with Conventional Airfoil Sections**, AIAA Journal, Vol. 18, Oct. 1980, pp. 1159-1164.

11. Boppe, C.W., **Transonic Flow Field Analysis for Wing-Fuselage Configurations**, NASA CR-3243, May 1980.

12. Tseng, K. and Morino, L., **Nonlinear Green's Function Method for Unsteady Transonic Flows**, Transonic Aerodynamics, Vol. 81 of Progress in Astronautics and Aeronautics, 1982.

13. Morino, L., **Steady, Oscillatory, and Unsteady Subsonic and Supersonic Aerodynamics - Production Version 1.1 (SOUSSA-P 1.1). Vol. 1, Thoeretical Manual**, NASA CR-159130, 1980.

14. Hessenius, K.A. and Goorjian, P.M., **Validation of LTRAN2-HI by Comparison with Unsteady Transonic Experiment**, AIAA Journal, Vol. 20, May 1982, pp. 731-732.

15. Yang, T.Y., Guruswamy, P., Striz, A.G., and Olsen, J.J., **Flutter Analysis of a NACA 64A006 Airfoil in Small Disturbance Transonic Flow**, Journal of Aircraft, Vol. 17, April 1980, pp. 225-232.

16. Yang, T.Y., Striz, A.G., and Guruswamy, P., **Flutter Analysis of MBB A-3 Supercritical Airfoil in Small Disturbance Transonic Flow**, Journal of Aircraft, Vol. 18, Oct. 1981, pp. 887-890.

17. Yang, T.Y., Guruswamy, P., and Striz, A.G., **Application of Transonic Codes to Flutter Analysis of Conventional and Supercritical Airfoils**, Journal of Aircraft, Vol. 19, March 1982, pp. 211-220.

18. Houwink, R. and van der Vooren, J., **Improved Version of LTRAN2 for Unsteady Transonic Flow Computations**, AIAA Journal, Vol. 18, Aug. 1980, pp. 1008-1010.

19. Edwards, J. W., Bennett, R.M., Whitlow, W. Jr., and Seidel, D.A., **Time-Marching Transonic Flutter Solution Including Angle-of-Attack Effects**, AIAA Paper 82-0685, New Orleans, Louis., May 1982.

20. Rizzetta, D. P. and Chin, W. C., **Effect of Frequency in Unsteady Transonic Flow**, AIAA Journal, Vol. 17, July 1979, pp. 779-781.

21. Isogai, K., **Transonic Dip Mechanism of Flutter of a Sweptback Wing. Part II**, AIAA Journal, Vol. 19, Sept. 1981, pp. 1240-1242.

22. Davis, S.S. and Malcolm, G.N., **Experimental Unsteady Aerodynamics of Conventional and Supercritical Airfoils**, NASA TM-81221, August 1980.

23. Tijdeman, H., van Nunen, J.W.G., Kraan, A.N., Persoon, A.J., and Poestkoke, R., **Transonic Wind Tunnel Tests on an Oscillating Wing with External Stores. Parts I-IV**, AFFDL-TR-78-194, Parts I to IV, Air Force Flight Dynamics Lab., 1978-1979.

24. Försching, H.W., **Grundlagen der Aeroelastik**, Berlin-Heidelberg-New York, Springer, Verlag, 1974.

25. Bisplinghoff, R.L., Ashley, H., and Halfman, R.L., **Aeroelasticity**, Reading, Mass., Addison-Wesley Publishing Company, 1955.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

A DYNAMIC MODEL OF ACCELERATION STRESS PROTECTION IN THE HUMAN AIRCREW MEMBER

Prepared by:	Patrick J. Sweeney, Ph.D., PE
Academic Rank:	Associate Professor
Department and University:	University of Dayton
	and
Prepared by:	Mary A. Doddy, M.S.
Academic Rank:	Graduate Student
Department and University:	University of Dayton
Research Location:	Air Force Aero Medical Research Laboratory Wright-Patterson AFB, Ohio
USAF Research Colleague:	Dr. Dana B. Rogers, Ph.D.
Date:	23 July 1982
Contract No.:	F49620-82-C-0035

A DYNAMIC MODEL OF ACCELERATION STRESS PROTECTION IN THE HUMAN AIRCREW MEMBER

by

Patrick J. Sweeney, Ph.D., P.E.

Mary Doddy, M.S.

ABSTRACT

This dynamic simulation computer model demonstrates the affects of G-forces upon the eyeball pressure of the aircrew member in various G-force environments. The stress relieving affects of G-suits and aircrew straining (M-1 and L-1 maneuvers) are dramatically depicted. The changes in the straining maneuver programs are evident in eyeball pressure and energy drain on the aircrew members. Further research in this area will undoubtedly result in improved aircrew performance in high and varying G-force environments.

ACKNOWLEDGMENTS

The authors wish to acknowledge the excellent and professional support of the entire Aero Medical Research Laboratory for their contribution in this effort.

Col. Mohr, AMRL Commander, very graciously explained the AMRL role in DOD activities and our part in this important mission. Ms. Patricia Lewandowski provided all of the necessary administrative support. Dr. Dana Rogers provided the authors with all the necessary data, background, and expertise for this report. He encouraged academic freedom and mission accomplishment during all of our many interactions. His study referenced in Section III was the major support study for this report.

The SCEEE staff and in particular Professor Peck and Dr. Steele were professional and supportive throughout this entire effort.

We thank all of these people, for without them the study could not have been accomplished.

A DYNAMIC MODEL OF ACCELERATION STRESS PROTECTION IN THE HUMAN AIRCREW MEMBER

I. INTRODUCTION:

Increased gravitational force results in a number of physiologic changes in the human exposed to such an acceleration. Classical methods of measuring these effects have led to an ever-increasing knowledge of the stressed physiologic systems. The two systems which have classically provided the defining limits in the stressed condition are vision and fatigue which then impinge on the performance capabilities of the human. Various performance measures have also been used to develop a greater knowledge of the acceleration environment. A multitude of model structures have arisen to describe the changes in oxygen levels, blood pressure, and in visual capability in hope of providing better descriptions and directions for advanced research. The model in this report has been developed from an alternative standpoint. The supposition is that there are competing systems within the human physiology which require oxygen in order to function at an acceptable level. The vision system requires a continuous flow of oxygen to the retinal layer to maintain function. The muscle system requires oxygen for replenishment of anaerobic activity and supply for aerobic activity. The central nervous system requires the same oxygen for its sustained higher level activity. The model in this development then becomes a model related to oxygen transport and competing energy used in the stressed physiologic system.

The total system of physiologic models was developed into a dynamic model according to the principles of system dynamics. System dynamics focuses on the structure and behavior of systems composed of interacting feedback loops. The resulting model integration allows investigation of the model in terms of an optimally controlled system. For example, varying acceleration time profiles can be run on the model resulting in unique physiologic response patterns for each profile. Performance is included through the pattern analysis inherent in the visual modulation transfer function and the cerebral oxygenation

prediction. The system model provides the basis for redefinition and extension of acceleration research into familiar areas with new intent.

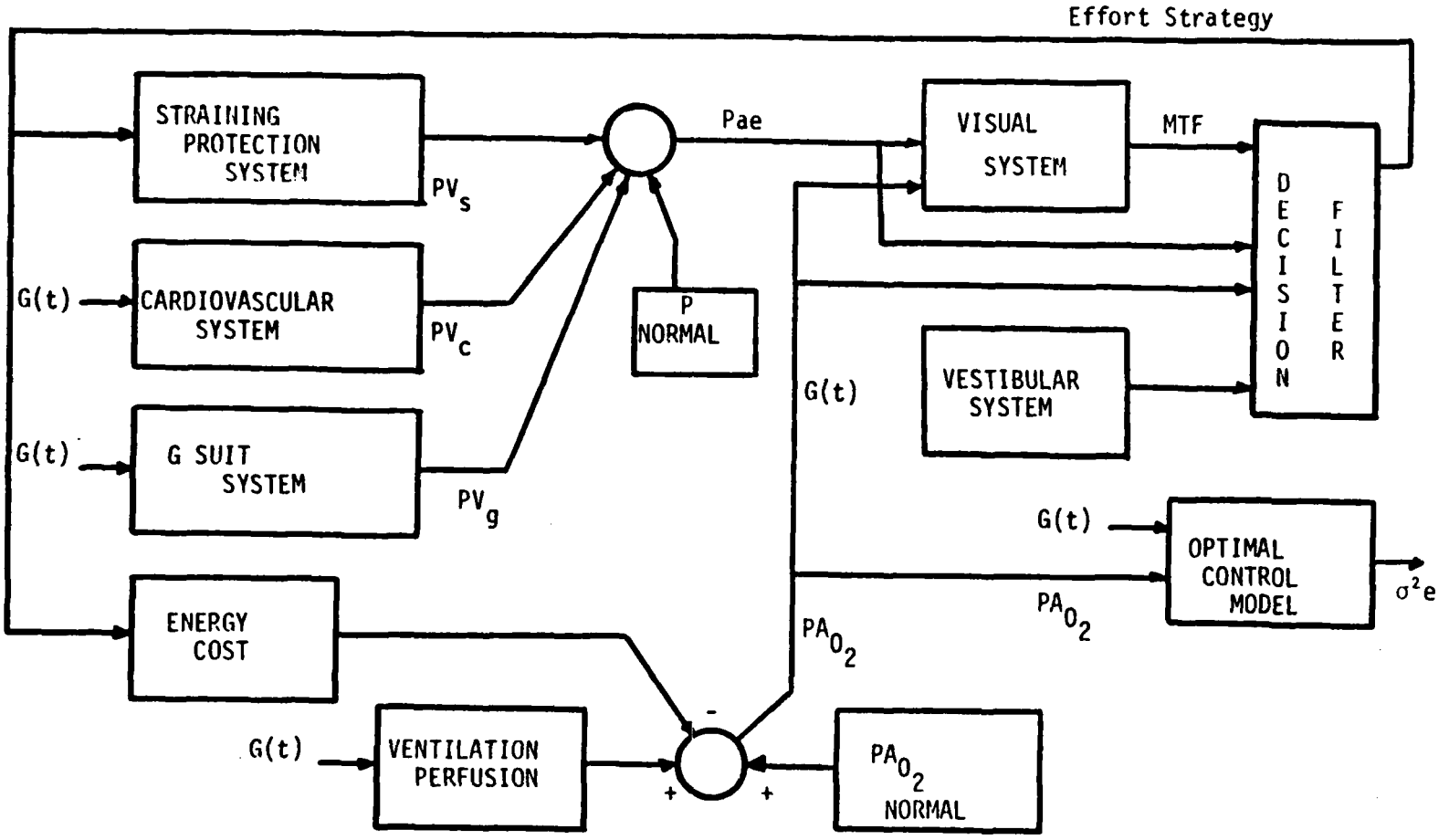
II. OBJECTIVES:

The main objective of this project was to dynamically model the acceleration stress protection in the human. This involved refining past models and submodels which were based largely on the experimental data available in the literature. The difficulties with the previous models are in the assumptions that each of the submodels is a linear system and does not have a cross-coupling affect with the other subsystems. By refining these models it was hoped that these linkages and couplings could be closely studied. The ultimate goal in refining these models was to arrive at a model that could be used as an optimal protection model.

III. MODEL:

A. Previously Developed Physiologic Models. Increased gravitational force stresses the human physiology in a manner that causes detrimental changes in the human's capability to maintain acceptable performance levels. A set of physiologic models which dynamically describe acceleration stress responses have been developed previously. Each model represents a major physiologic system or acceleration protection mechanism. The submodels are assembled as a larger system using natural linkage variables from one subsystem to the next. The major subsystems are the visual system, cardiovascular subsystem, G-suit system, and the energy cost system. The integrated system model links all the subsystems as shown in Figure 1.

Human tolerance to long-term $+G_z$ acceleration is normally measured in terms of visual loss (blackout) and unconsciousness. Both of these tolerance end points are related to the ability of the cardiovascular system to deliver oxygenated blood at adequate pressure to the retinal and cerebral regions. The acceleration causes a changing blood pressure profile in the human such that the effective pressure vertically above the heart level is decreased and the pressure below the heart level is increased. There is therefore a lower perfusion pressure at



LEGEND

E(t) = Effort Strategy

G(t) = Acceleration Vector

PVs = Protection Value of Straining

PVc = Protection Value of Cardiovascular

PVg = Protection Value of G Suit

Pae = Arterial Pressure at eye level

PA_{O_2} = Arterial Oxygen Saturation in %

$\sigma^2 e$ = General Performance Metric

MTF = Modulation Transfer Function

eye level. The distribution of the blood in the body also changes as the acceleration pools blood in the lower part of the body and lungs. There is therefore less available blood to circulate and a lower oxygen content because the lungs do not operate as efficiently. The cardiovascular response model output provides a dynamically responding signal which is equivalent to nominal eye level blood pressure values for an unprotected human undergoing the equivalent acceleration.

The effect that acceleration has on the visual apparatus is observed in terms of tunnel vision, greyout, and blackout. During the periods of greyout there are also decreases in visual acuity and brightness contrast detection ability. Although there are multiple factors related to the anatomy, psychology, and physiology of the human which are responsible for these changes in visual perception.

The structure of the eye is a primary factor and provides the basis for a usable model. When the blood pressure supply drops below a critical point the retinal supply vessels no longer remain open to blood flow. The changing pressure levels which are postulated across the eye serve as a partial explanation and a functional model for the prediction of field changes other than total blackout. The dynamic visual field model reacts to G level inputs from an external source and produces a dynamically responsive signal which predicts the expected visual field of a pilot undergoing the identical G profile.

The G suit model provides the parameters which describe the valve system and the garment used to increase the individual's tolerance to $+G_z$. The suit uses pressurized bladders to press against the legs and lower abdomen. The external pressure inhibits displacement of the blood volume to the lower extremities thus insuring a better blood supply to the heart during acceleration. The suit must be inflated by the G valve to a predetermined level for the suit to be effective. The simulation model accounts for the required pressure level and uses the actual suit to provide the necessary dynamics. The suit pressure is compared with the required schedule and a protection value is generated by the model.

The ventilation perfusion model allows the effects of reduced perfusion to be incorporated. Acceleration induced shifts in the

pulmonary blood supply reduce the amount of oxygen transport across the alvedlar surface and thus reduce the oxygen supply.

The straining system model accounts for the G tolerance enhancement which is afforded by a properly executed M-1 or L-1 maneuver. The purpose of these straining maneuvers is to increase the blood pressure delivered to the eye. Proper performance of the maneuvers requires that the abdominal and upper torso muscles be tensed isometrically and that expirations should be made against a partially closed or closed glottis. The result is an increased intrathoracic pressure and increased blood pressure at the eye.

The energy stores/oxygen model is an energy balance system in which oxygen is proficed through the pulmonary system. The oxygen is utilized with the metabolic fuels in the straining system with the resultant higher blood pressure through the straining. Other oxygen is utilized in normal metabolic maintenance and the oxygen reserves are depleted. The energy expenditure model is simply an integrator which integrates the energy effort and translates this to a depletion of the oxygen stores in terms of a decreased percent of oxygen.

B. The Computer Model. The computer simulation model was based upon the 1979 study entitled "A Model for the Energetic [illegible]t of Acceleration Stress Protection in the Human" by Dr. Dana B. Rogers. This report concluded by stating that "the use of dynamic models . . . allows for a variety of speculative studies." This report is a dynamic model that provides the vehicle necessary to evaluate the effects of varying acceleration loads on aircrew members.

The simulation language used in this study was developed by Dr. Jay Forrester of MIT. The language is DYNAMO and is based upon feedback interactions. Important variables are measured as levels (L on the right-hand margin in the model). The change factors are rates (R) and affect the inflow and outflow to the levels. These rates are affected by auxiliary (A) equations, constants (C), and initial conditions (N). In addition, table functions (T) are included when the input is a function of a specific variable.

```
0. L  PAE.K=MIN(80,PAE.J+DT(ETRO.JK+PVG.JK-PVC.JK))
0. N  PAE=79
2. *   PAE=PRESSURE AT THE EYEBALL
3. *   PVG=PRESSURE INCREASE DUE TO G SUIT
4. *   ETRO=ENERGY DRAIN
5. *   PVC=PRESSURE DUE TO G FORCES ALONE
6. R  PVC.KL=(CVS.K*FAC)
7. *   FAC=CONSTANT FOR CVS
8. *   CVS=G FORCE SLOPE (RATE OF CHANGE)
9. C   FAC=5
```

Equation 1 is a level equation that measures the blood pressure at the eyeball. Lower pressures reduce initially peripheral vision and at higher G-forces reduces visual to "tunnel vision" and subsequently to blackout. This eyeball pressure (PAE) is affected by the energy level (ETRO) of the aircrew members, the G-suit effects (PVG) and the G-forces encountered by the aircrewmen (PVC).

CVS is the actual G-force rate of change pattern on the crew member and the rate of change in G-force in PVC. For scaling and dimensional purity, FAC is included as a constant in the PVC equation.

```
10. A  CVS.K=TABHL(CVST,TIME.K,0,392.5,.5)
11. T  CVST=0,1,1,1,1,
12. X  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
13. X  1,1,1,1,1,1,1,1,1,
14. X  0,0,0,0,0,0,0,0,0,0,0,
15. X  -1,-1,-1,-1,-1,
16. X  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
17. X  -1,-1,-1,
18. X  0,0,0,0,0,0,0,0,0,0,0,
19. X  1,1,1,1,
20. X  0,0,0,0,0,0,0,0,0,0,0,
21. X  -1,-1,
22. X  0,0,0,0,0,0,0,0,0,0,0,
23. X  1,1,1,1,1,1,1,
24. X  0,0,0,0,0,0,0,0,0,0,0,
25. X  -1,-1,-1,-1,-1,-1,-1,-1,-1,
26. X  0,0,0,0,0,0,
27. X  1,1,1,1,1,1,1,
28. X  0,0,0,0,0,0,0,0,0,0,0,
29. X  -1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,
30. X  1,1,1,1,1,1,
31. X  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
32. X  0,1,1,1,1,1,1,1,1,
33. X  0,0,0,0,0,0,0,0,0,0,0,
34. X  -1,-1,-1,-1,-1,
```

```
35. X  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
36. X  -1,-1,-1,
37. X  0,0,0,0,0,0,0,0,0,0,0,
38. X  1,1,1,1,
39. X  0,0,0,0,0,0,0,0,0,0,
40. X  -1,-1,
41. X  0,0,0,0,0,0,0,0,0,0,
42. X  1,1,1,1,1,1,
43. X  0,0,0,0,0,0,0,0,0,0,
44. X  -1,-1,-1,-1,-1,-1,-1,-1,
45. X  0,0,0,0,0,
46. X  1,1,1,1,1,1,
47. X  0,0,0,0,0,0,0,0,0,0,
48. X  -1,-1,-1,-1,-1,-1,-1,-1,-1,-1,
49. X  1,1,1,1,1,
50. X  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
51. X  0,1,1,1,1,1,1,1,
52. X  0,0,0,0,0,0,0,0,0,0,
53. X  -1,-1,-1,-1,-1,
54. X  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
55. X  -1,-1,-1,
56. X  0,0,0,0,0,0,0,0,0,0,
57. X  1,1,1,1,
58. X  0,0,0,0,0,0,0,0,0,0,
59. X  -1,-1,
60. X  0,0,0,0,0,0,0,0,0,0,
61. X  1,1,1,1,1,1,
62. X  0,0,0,0,0,0,0,0,0,0,
63. X  -1,-1,-1,-1,-1,-1,-1,-1,
64. X  0,0,0,0,0,
65. X  1,1,1,1,1,1,
66. X  0,0,0,0,0,0,0,0,0,0,
67. X  -1,-1,-1,-1,-1,-1,-1,-1,-1,-1,
68. X  1,1,1,1,1,
69. X  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
70. X  0,1,1,1,1,1,1,1,
71. X  0,0,0,0,0,0,0,0,0,0,
72. X  -1,-1,-1,-1,-1,
73. X  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
74. X  -1,-1,-1,
75. X  0,0,0,0,0,0,0,0,0,0,
76. X  1,1,1,1,
77. X  0,0,0,0,0,0,0,0,0,0,
78. X  -1,-1,
79. X  0,0,0,0,0,0,0,0,0,0,
80. X  1,1,1,1,1,1,
81. X  0,0,0,0,0,0,0,0,0,0,
82. X  -1,-1,-1,-1,-1,-1,-1,-1,
```

```
 83. X  0,0,0,0,0,
 84. X  1,1,1,1,1,1,
 85. X  0,0,0,0,0,0,0,0,0,0,
 86. X  -1,-1,-1,-1,-1,-1,-1,-1,-1,-1,
 87. X  1,1,1,1,1,
 88. X  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
 89. X  0,1,1,1,1,1,1,1,
 90. X  0,0,0,0,0,0,0,0,0,0,
 91. X  -1,-1,-1,-1,-1,
 92. X  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
 93. X  -1,-1,-1,
 94. X  0,0,0,0,0,0,0,0,0,0,
 95. X  1,1,1,1,
 96. X  0,0,0,0,0,0,0,0,0,0,
 97. X  -1,-1,
 98. X  0,0,0,0,0,0,0,0,0,0,
 99. X  1,1,1,1,1,1,
100. X  0,0,0,0,0,0,0,0,0,0,
101. X  -1,-1,-1,-1,-1,-1,-1,-1,
102. X  0,0,0,0,0,
103. X  1,1,1,1,1,1,
104. X  0,0,0,0,0,0,0,0,0,0,
105. X  -1,-1,-1,-1,-1,-1,-1,-1,-1,-1,
```

As stated previously, CVS is the actual G-force rate of change exerted on the aircrewman. This G-profile has been provided by AMRL and was replicated five times to reflect a 400 second G-pattern. This force is depicted as the derivative of the actual G-force, because of the cumulative nature of the level equation, PAE, and the appropriate rate, PVC.

```
106. A  PVGD.K=DELAY1(CVR.K*FAC1,1)
107. *   PVGD=G SUIT PRESSURE DELAY FUNCTION
108. *   CVR=G SUIT-G FORCE FACTOR
109. *   FAC1=G SUIT-G FORCE CORRECTION FACTOR
110. *   FAC1=0.5
```

PVGD is an auxiliary equation that accounts for the effects of the programmed G-program and the G-suit system delay. G-suits increase the pressure to the eyeball but are affected by mechanical system operational delays. The delay is a single period (0.1 second delay) and is represented by a 1 in equation 106. FAC1 is a sizing and dimensional constant.

```
111. A  CVR.K=CVS.K*FACX
112. *   FACX=CVS TO CVR CORRECTION FACTOR
113. C  FACX=1
114. R  CVX.KL=CVR.K
115. *   CVX=G SUIT-G FORCE RATE FACTOR
116. L  CXX.K=CXX.J+DT(CVX.JK)
117. *   CXX=APPLIED G FORCE
118. N  CXX=1
```

CVR is a correction factor used to convert CVS to usable form in equation 106. FACX is the dimensional constant.

CVX is the equation used to convert CVR to a rate.

CXX is a level equation that is used to show the actual G-force on the aircrewman. This along with PAE, ET, and other important factors are plotted in the ensuing figure.

```
119. L  ACXX.K=ACXX.J+(DT/AT)(CXX.J-ACXX.J)
120. *   ACXX=AVERAGE APPLIED G FORCE
121. *   AT=AVERAGING DELAY FUNCTION
122. N  ACXX=0
123. C  AT=10
```

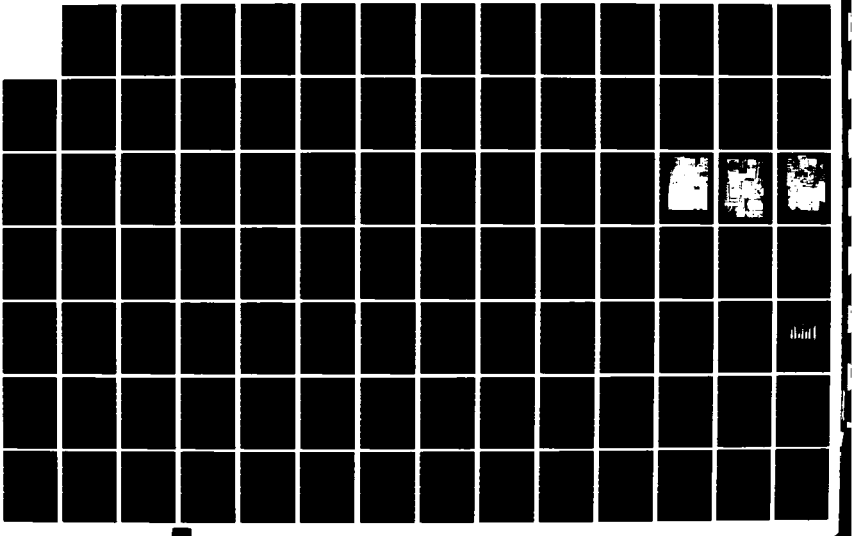
ACXX is the G-force averaged over the previous 10 periods. ACXX is initialized at 0, see line 122. AT is the averaging period of 10.

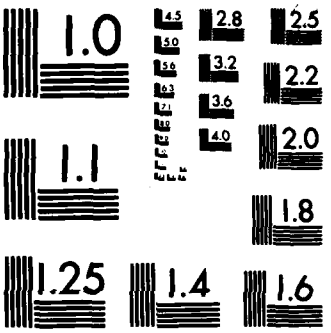
```
124. R  PVG.KL=MAX(CLIP(PVGD.K,0,80,PAE.K),0)*FACET
125. *   PVG=PRESSURE INCREASE DUE TO G SUIT
126. C  FACET=2
127. *   ENERGY LEVEL - G-SUIT PRESSURE LEVEL CONSTANT
```

PVG is the G-suit effect. Increasing G-forces will affect the G-suit input delayed by the system capabilities.

```
128. L  ET.K=ET.J+DT(ETRI.JK-ETRO.JK)*FAC10
129. C  FAC10=2
130. N  ET=100
131. R  ETRI.KL=ETRIC
132. C  ETRIC=0
133. R  ETRO.KL=CLIP(ETD.K/FAC2,0,80,PAE.K)*FAC8
134. C  FAC8=.5
```

AD-A130 770 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM (1982) 9/11
RESEARCH REPORTS VOLUME.. (U) SOUTHEASTERN CENTER FOR
ELECTRICAL ENGINEERING EDUCATION INC S..
UNCLASSIFIED W D PEELE ET AL. OCT 82 AFOSR-TR-83-0614 F/G 5/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

```
135. C  FAC2=10
136. *   ET=ENERGY LEVEL OF THE AIRCREWMAN
137. *   ETRO=ENERGY OUTPUT RATE
138. *   ETRO=ENERGY INPUT RATE
139. *   ETRIC=ENERGY INPUT RATE CONSTANT
140. *   ETD=PILOT ENERGY EXPENDITURE PROGRAM
141. *   FAC2=ENERGY EXPENDITURE CONSTANT
```

ET is the energy level of the aircrew member. As energy is expended (ETRO) the energy level decreases from 100 percent. This would be restored to 100 percent when the G-forces were reduced to 1 over an extended period of time; however, in this model ETRI (Energy input rate is equal to years), since the G-program extends for 400 seconds. Both FAC10 and FAC8 are dimensional constants.

```
142. A  ETD.K=TABHL(ETDT,TIME.K,0,400,5)
143. T  ETDT=0,5,5,0,0,5,5,0,0,5,5,0,0,5,5,0,0,5,5,0,0,5,5,0,
144. X  0,5,5,0,0,5,5,0,0,5,5,0,0,5,5,0,0,5,5,0,0,5,5,0,0,5,5,0,
145. X  0,5,5,0,0,5,5,0,0,5,5,0,0,5,5,0,0,5,5,0,0,5,5,0,0,5,5,0,0
```

ETD is a table function that reflects the M-1 or L-1 maneuver and the quantity of energy expended by the aircrew member. The 400 and 5 in equation 142 shows that in equations 143-145 the effort extends for 400 seconds and the values are 5 seconds apart.

```
146. A  PAO2.K=(A*ACXX.K-B*CXX.K*CXX.K)(D-EXP((E*TIME.K)*LOGN(F)))+G
147. *   PAO2=PERCENT OXYGEN SATURATION
148. C  A=.5
149. C  B=.4
150. C  D=1
151. C  E=-.05
152. C  F=2.71828
153. C  G=98
154. *   A=CONSTANT
155. *   B=CONSTANT
156. *   D=CONSTANT
157. *   E=CONSTANT
158. *   F=CONSTANT(E)
159. *   G=CONSTANT(MAXIMUM PAO2 SATURATION)
```

PAO2 is the percent oxygen saturation of the aircrewman. This equation was taken directly from the Rogers study and mathematically is:

$$PAO2 = H(.5g-.4g^2) \times (1-e^{-.05t}) + 98$$

Constants A-G are noted in equations 148-153.

```
160. A  EVP.K=ET.K*PAO2.K*H
161. C  H=.0001
162. A  VS.K=PAE.K*EVP.K*S/R
163. C  R=79
164. C  S=1
165. *   EVP=ENERGY OXYGEN FUNCTION
166. *   H=CONSTANT
167. *   VS=EYEBALL PRESSURE AND ENERGY-OXYGEN FUNCTION
168. *   R=CONSTANT
169. *   S=CONSTANT
```

EVP is the energy-oxygen relationship and reflects that energy is depleted faster when the oxygen supply is reduced. H is the dimensional constant.

VS is a measure of available field of vision as a function of eyeball pressure and the energy-oxygen function. R and S are dimensional constants. Note: Since some argument does exist concerning visual field and decision-making capability this factor was deemed beyond the scope of this ten-week study.

```
170. C  PRTPER=0
171. PRINT PAE,CVS,PVC,PVG,ETRO,CXX,ET
172. C  DT=.1
173. C  LENGTH=400
174. C  PLTPER=5
175. PLOT PAE=P/PVC=V/ETRO=E/PVG=G/CXX=X/ET=%/PAO2=*VS=S
```

The above functions control the operation of the model and the output. PRTPER is the printer command and controls how often a PRINT item is printed. PRINT controls what is printed. DT is the rate at which the DYNAMO compiler computes. LENGTH is the length of the printout. PLTPER defines how often the PLOT items are plotted. PLOT determines what is plotted.

IV. RESULTS AND CONCLUSIONS:

The model reflects the affects of varying G-forces on an aircrewman in changes in eyeball pressure, oxygen level, and energy level. The cumulative affects of these variables alter visual acuity and subsequently the aircrew member's ability to make decisions and accomplish his immediate mission.

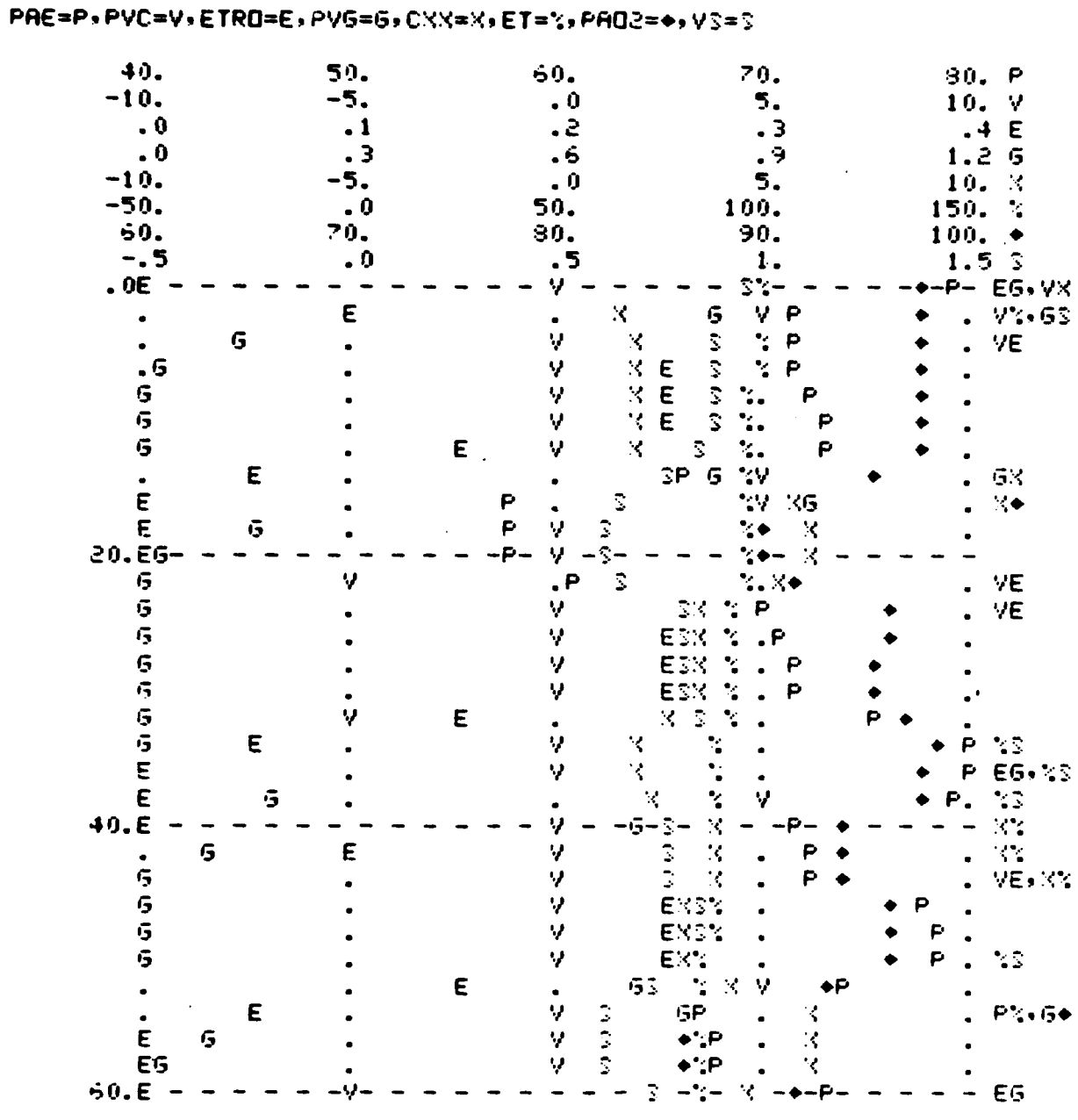
Throughout the model constant factors have been included so as to facilitate easy value changes for future work. However, as the model is currently structured reasonable results are created. In Figure 2, Model Output--Basic Run, the following variables are plotted:

PAE(P) - Pressure at the Eyeball
PVC(V) - Pressure due to G-force
ETRO(E) - Energy Drain in Pressure
PVG(G) - Pressure Increase Due to G-Suit
CXX(X) - Actual G-Force on the Aircrewman
ET(%) - Energy Level of the Aircrewman
PAO2(*) - Percent Oxygen Saturation
VS(S) - Eyeball Pressure and Oxygen Function

As G-forces increase PAE, PVC, ETRO, ET, PAO2, and VS, all respond in the proper directions and at what the authors believe to be in a reasonable range. As the G-pattern continues and the aircrewmember executes either M-1 and L-1 maneuvers he eventually depletes his energy level, ET, and is unable to continue with the pressure reducing maneuvers. He subsequently blacks out. The G-suit has a similar but opposite affect upon the model.

The sensitivity analysis shows the affects of variable constants and M-1 or L-1 maneuver patterns. As the aircrewman alters his maneuvering pressure level and/or pattern, the affects of these maneuvers upon his eyeball pressure changes appropriately. The model shows little difference when the aircrewman holds the M-1 or L-1 for "long periods" or whether he executes a large number by not holding for long periods on each individual maneuver.

The actual runs are not included due to the required limited brevity of this report.


PAE=P, PVC=V, ETRO=E, PVG=G, CXX=X, ET=%, PAO2=◆, VS=S
40. 50. 60. 70. 80. P
-10. -5. .0 5. 10. V
.0 .1 .2 .3 .4 E
.0 .3 .6 .9 1.2 G
-10. -5. .0 5. 10. X
-50. .0 50. 100. 150. %
60. 70. 80. 90. 100. ◆
-.5 .0 .5 1. 1.5 S
.0E V S% ◆ P EG,VX
V%,GS
VE
20.EG P
VE
VE
40.E V %
VE,X%
P%,G◆
60.E V P EG

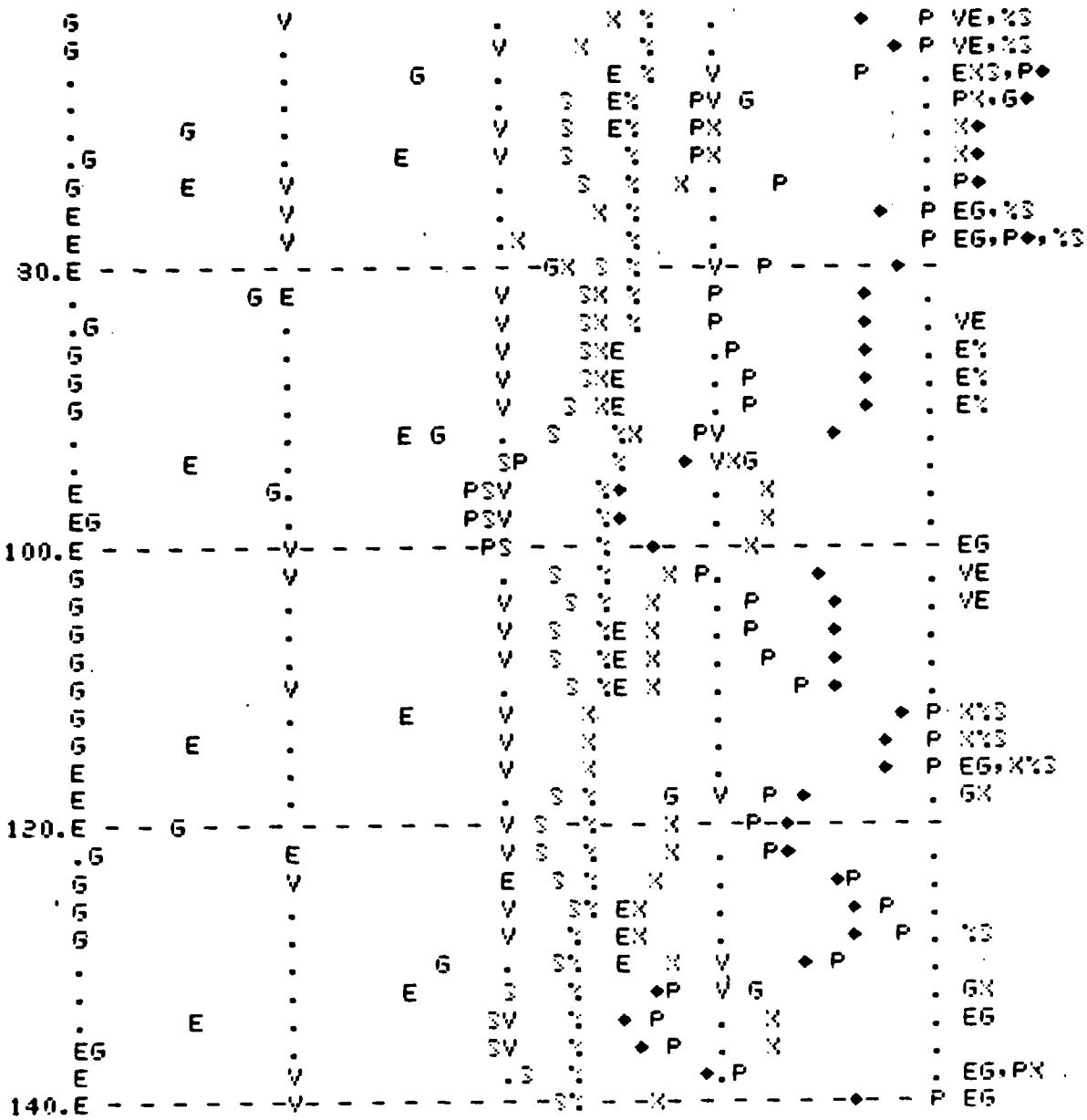
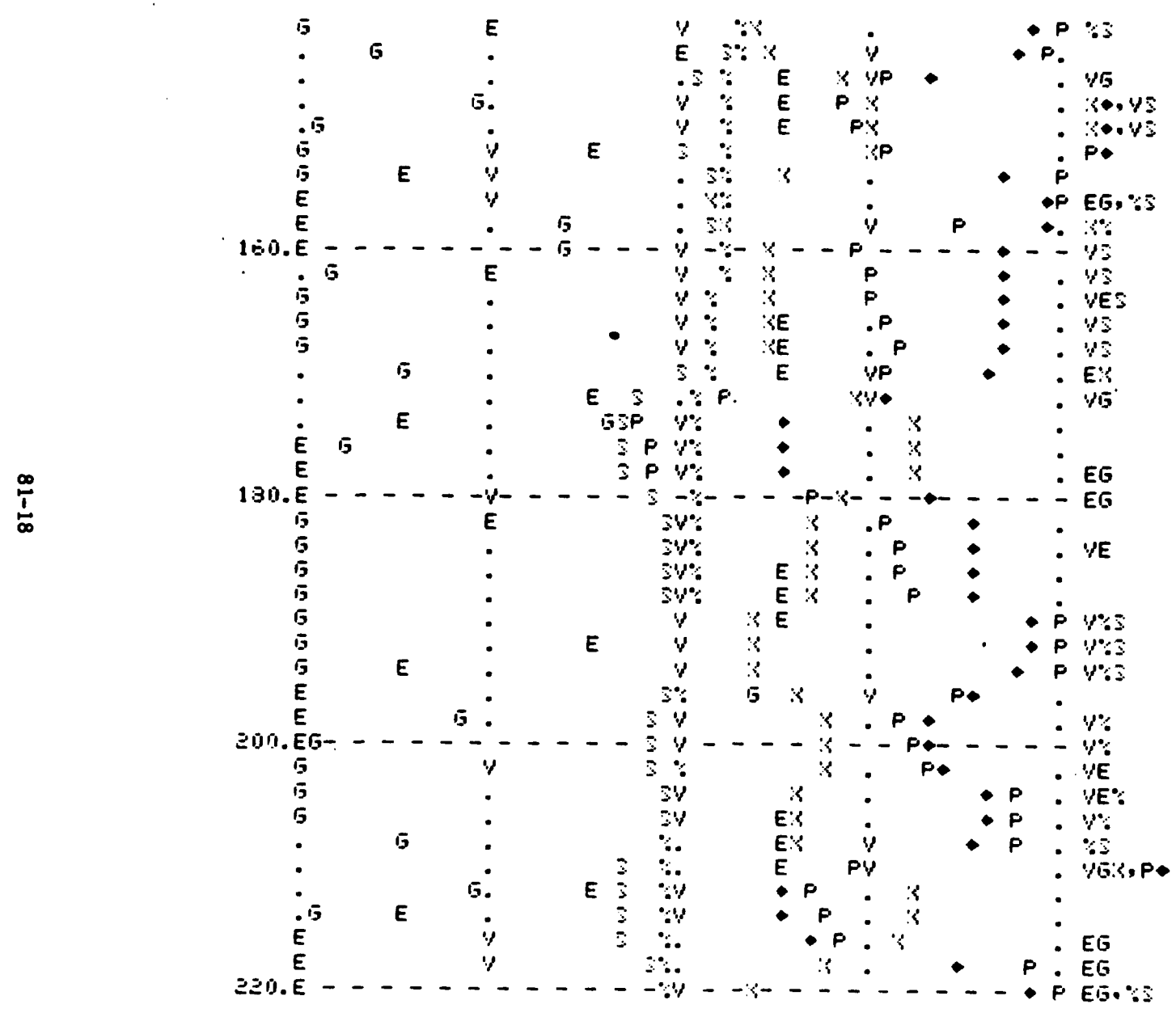


Figure 2--Continued



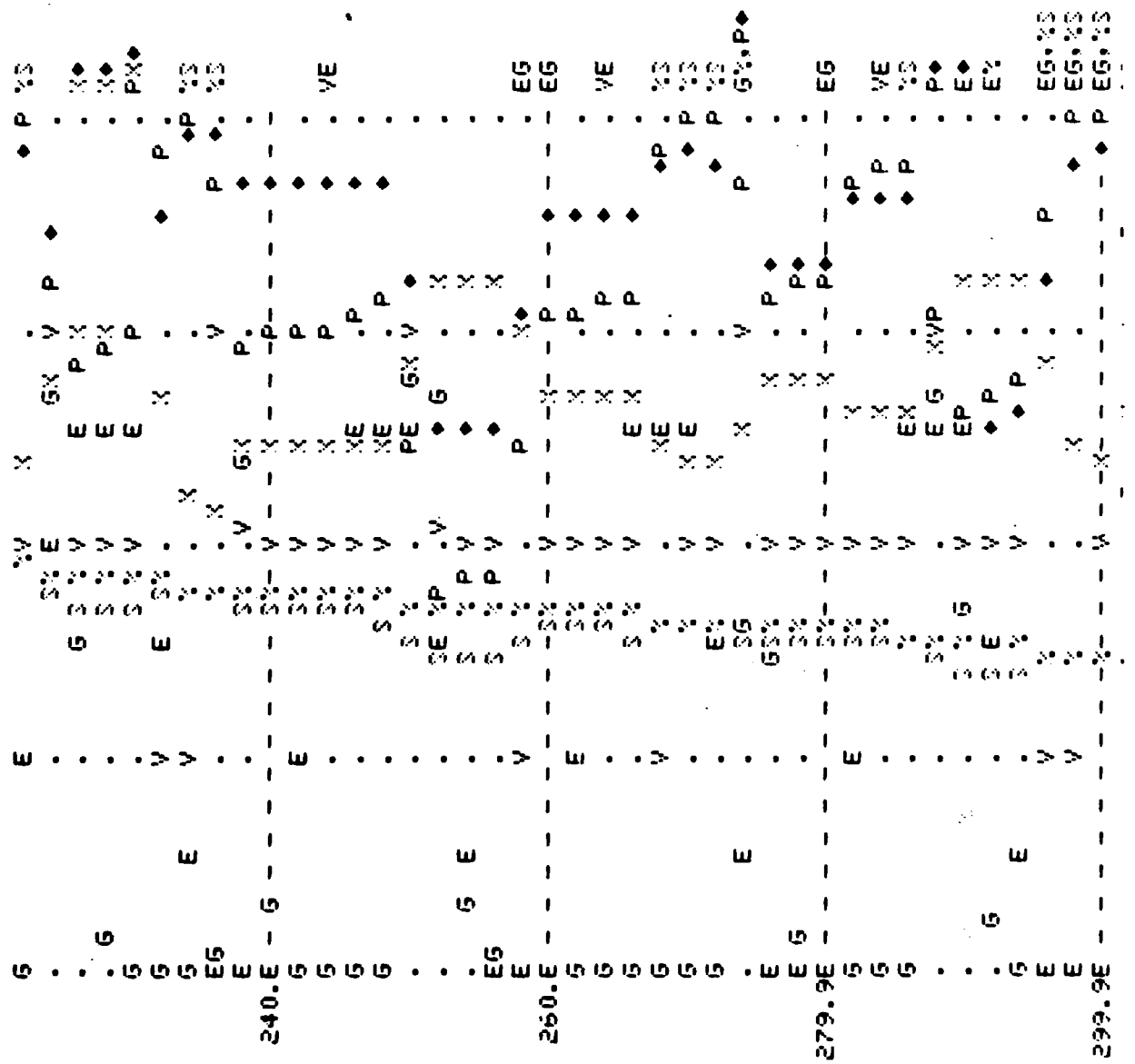


Figure 2--Continued

V. Recommendations

This type of study has great promise, for like the aircraft simulator, this model has the potential of testing without flying. The authors will apply for a mini-grant due to the high value potential of this work.

The authors suggest that the model be validated with actual test data. Modifications can be easily made since the model has captured the essence of the problem.

With the validated model the analyst can create enough scenarios and test them so as to develop the best M-1 or L-1 maneuver program based on the projected mission profile.

The authors believe that since each mission is plagued with its own unique G-pattern that the aircrewman should be provided with a suggested M-1 or L-1 program based upon the anticipated short and long term G-patterns of the mission.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

COLLISION-INDUCED ABSORPTION IN THE FAR INFRARED SPECTRUM OF N_2

Prepared by:	Dr. Richard H. Tipping
Academic Rank:	Professor
Department and University:	Department of Physics University of Nebraska at Omaha
Research Location:	Air Force Geophysics Laboratory Hanscom Air Force Base Bedford, MA 01731
USAF Research Colleague:	S.A. Clough
Date:	September 15, 1982
Contract No:	F49620-82-C-0035

COLLISION-INDUCED ABSORPTION IN THE FAR INFRARED SPECTRUM OF N_2

by

Richard H. Tipping

ABSTRACT

The far infrared absorption arising from collision-induced transitions in gaseous N_2 has been investigated. After a review of the procedure for the theoretical calculation of the translation-rotational spectrum, new results for a temperature of 200K are reported. Similar calculations for other temperatures appropriate to the Earth's atmosphere will enable one to model the absorption by N_2 for arbitrary paths through the stratosphere. The collision-induced absorption mechanism becomes important only for paths above the tropopause, where the absorption by allowed water transitions and by the corresponding water vapor continuum absorption that dominate far infrared spectra near the Earth's surface are small.

Acknowledgement

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for their financial assistance. He would also like to acknowledge the hospitality provided by the Air Force Geophysics Laboratory, Hanscom Air Force Base.

Finally, he would like to thank S.A. Clough, F.X. Kneizys, and W.O. Gallery for helpful discussions throughout the summer.

I. INTRODUCTION:

The absorption of microwave and infrared radiation by molecules in the Earth's amtosphere is a problem of considerable importance from both the theoretical and the practical standpoints[1]. In atmospheric windows (e.g., regions centered around 1000 cm^{-1}, 2500 cm^{-1}, etc.) and between allowed vibration-rotational transitions, there exists appreciable residual absorption. Near the Earth's surface, this "continuum absorption" is attributed to water molecules present in substantial numbers and extends from the microwave region to several thousand wave numbers[2]. In the spectral region above 2400 cm^{-1}, another broad, temperature-dependent absorption arising from the collision-induced fundamental vibration-rotational transitions in N_2 is observed in long path measurements through the atmosphere[3]. This type of absorption results from transient dipoles created and modulated by collisions. It has been observed in the laboratory[4,5] and is well understood theoretically[6]. A similar absorption by N_2 in the far infrared peaking around 100 cm^{-1} is associated with rotation-translational transitions. This would provide an important source of opacity in those regions in the atmosphere where the concentration of water molecules is low (e.g., above the tropopause). Because of the nature of the induction process, binary collisions predominate and the absorption coefficient is therefore proportional to the square of the N_2 density. However, in order to model this absorption in the atmosphere, one also needs to know how the absorption depends on temperature.

II. OBECTIVES OF THE RESEARCH EFFORT

The primary objective of this research was to synthesize the pure rotation-translational spectrum of N_2. In order to accomplish this, laboratory spectra had to be analyzed in terms of the molecular parameters and spectral line shapes that characterize the various induction mechanisms. With this information, the far infrared spectrum of molecular nitrogen can be generated for specified temperatures and densities. Eventually, the aim of the project is to incorporate this information into existing computer codes which enable one to calculate the absorption of radiation along arbitrary paths through the atmosphere.

III. COLLISION-INDUCED TRANSLATION-ROTATIONAL ABSORPTION IN N_2

In this section we describe the procedure that is used to generate the far infrared spectrum of pure nitrogen and present new quantitative results. The calculation is performed in terms of the quantity $G(\nu)$ which is defined by

$$A(\nu) = 1/2\ \rho^2\ \nu\ (1-e^{-\beta hc\nu})G(\nu) \quad . \qquad (1)$$

In this expression, $A(\nu)$ is the absorption coefficient

$$A(\nu) = \ell^{-1}\ \ln\ (I_o(\nu)/I(\nu)) \qquad (2)$$

where $I_o(\nu)$ and $I(\nu)$ are respectively the incident and transmitted intensities at frequency ν (cm^{-1}), and ℓ is the path length. The density of the nitrogen gas is denoted by ρ (amagat units), and $\beta = (kT)^{-1}$ where k is Boltzmann's constant and T is the temperature (Kelvin). The quantity $G(\nu)$ is taken as independent of the density and corresponds physically to the spectral density associated with a pair of molecules.

Expression (1) is therefore valid only for densities and frequencies at which binary interactions are dominant. It has been shown by Bosomworth and Gush[5] that this is the case for nitrogen densities at least up to 35 amagat, and thus this formalism is certainly adequate for conditions prevailing in the atmosphere.

The dipole moments responsible for the absorption are induced dipoles which are created as two N_2 molecules approach each other during a collision. The main contribution to the induced dipole is due to the polarization of one of the molecules by the quadrupolar electric field of the other. This dipole has the following form

$$\vec{\mu} = \alpha_1 \vec{E}_2 + \alpha_2 \vec{E}_1 \qquad (3)$$

where α_1 and α_2 are the isotropic polarizabilities and $\vec{E}_1$ and $\vec{E}_2$ are the electric quadrupolar fields of molecules 1 and 2, respectively. The quadrupolar field of a molecule (apart from angular factors) is proportional to R^{-4}, where R is the separation between the centers of mass of the two molecules, and to the quadrupole moment Q. This mechanism gives rise to absorption with selection rules of $\Delta J = 0, \pm 2$, where J denotes the rotational quantum number of an isolated molecule. Well over 90% of the total absorption in the far infrared band is due to this induction mechanism. This dipole (apart from angular factors) and the characteristics of the transitions to which it gives rise, are summarized in the first line of Table I. The rest of the absorption is associated with various other mechanisms for creating induced dipoles. For example, there is a dipole analogous to the one discussed above but associated with the anisotropic part, γ, of the polarizability

tensor. Apart from the angular factors, this dipole is proportional to $\gamma Q/R^4$ and the corresponding selection rules and transitions are given in the second line of Table I. Note that this mechanism allows for so-called "double transitions" in which both molecules change their rotational state in the absorption process; these are indicated by the notation: S(J)+S(J'), etc.

In addition, there are induced dipoles due to the next higher multipole of N_2, viz. the hexadecapole moment Φ. Even though the contribution of this hexadecapolar induction mechanism is only of the order of 5% of the total integrated intensity, it turns out that this is the dominant mechanism in the higher frequency part of the spectrum ($\nu > 100$ cm^{-1}). The reason for this is that the hexadecapole-induced absorption satisfies the selection rules $\Delta J = 0, \pm 2, \pm 4$, and, therefore, some of the corresponding transitions occur at higher frequencies. These dipoles are proportional to $\alpha\Phi/R^6$, and the characteristics of these transitions are listed in line 3 of Table I. The analogous contributions arising from the anisotropy of the polarizability are characterized in line 4.

Another mechanism for creating transient dipoles is associated with the overlap interaction between the colliding molecules, which is not described in terms of multipolar fields as the ones above. The most important so-called overlap contribution to the dipole is the one that describes the deviation of the dipole given by Eq. (3) from its R^{-4} dependence. This mechanism is called the "anisotropic L=3" overlap; it is characterized by a strength parameter λ_3 and a range parameter ρ_3. This gives rise to the same transitions as the isotropic

quadrupolar induction and is therefore also listed in line 1 of Table I. Because these two mechanisms always occur together and have the same angular dependencies, an interference between them also occurs. The final overlap mechanism which we include is the "anisotropic L=1" component; this has parameters λ_1, and ρ_1, and is characterized on line 5 of Table I. It turned out that for a detailed description of the spectrum in the temperature range from 124 to 300K, all of the above mechanisms are necessary and thus have been included in our present work.

It can be shown that, in the approximation of an isotropic intermolecular interaction, the spectrum consists of a sum of lines, each associated with an operator listed in Table I; we will use this approximation in the generation of the total spectrum. Each of the contributing lines has its own characteristic shape and width which are determined for the most part by the R-dependence of the induced dipole. For example, the short-range overlap mechanism gives rise to broad lines, while the long-range quadrupolar induction leads to much narrower lines.

Bosomworth and Gush[5] have shown that the quadrupolar lines of H_2 are well represented by a lineshape which consists of a Lorentzian with an exponential tail attached at a point $p\delta$, where δ is the width (half-width at half-maximum) parameter of the Lorentzian and p is a temperature dependent parameter of the order of 1 to 3. In addition, the profile is multiplied by a factor in order to satisfy the detailed balance condition. We have therefore used this "Lorentz-Exponential" (L-E) line shape model to represent all of the transitions of nitrogen

except for the anisotropic L=1 components. The L-E profile is characterized by three parameters: an intensity parameter, and p and δ discussed above. Two of these three parameters can be determined (in terms of the multipole moments, polarizabilities, etc.) by requiring the profile to satisfy the exact theoretical values for the first two spectral moments[7]. We determine the intensity parameter and p in this way, and adjust δ until agreement is reached between the synthetic spectrum and the experimental one[5]. The δ parameters for the different mechanisms are given in Table I for T=300K. As is well known , the widths scale approximately as $T^{1/2}$.

The lines associated with the anisotropic L=1 mechanism can be represented by a two parameter line shape involving a modified K_2 Bessel function[8]. These two parameters are uniquely determined by requiring this profile to satisfy the first two moment relations also.

In computing the values for the first two moments of each line, we use the isotropic part of the intermolecular potential given by Billing and Fisher[9] in order to represent the pair distribution function. For the temperature range of interest, the translational motion of the N_2 molecules can still be treated classically, and the required radial integrals have been carried out numerically. The resulting molecular parameters that characterize the spectrum for the temperature range from 124 to 300K are given in Table II; these parameters should also be appropriate for the temperatures of interest in the present study.

Once the line profile models are established, the total translation-rotational spectrum is built up by adding together all of

the lines (with appropriate values of δ) that the various induction mechanisms produce. The number of rotational states that contribute significantly to the synthetic spectrum depends, of course, on the temperature; it varies from approximately 15 at 70K to nearly 30 at 300K.

Using the procedure discussed above, we have synthesized the N_2 spectrum from 0 to 600 cm^{-1}. Over this range of frequencies, the functions $G(\nu)$ and $A(\nu)$ vary by many orders of magnitude; this is illustrated in Table III in which we list $G(\nu)$ versus ν for a temperature of 200K. Similar results for other temperatures can also be generated.

IV. RECOMMENDATIONS

As a result of the current research, it is recommended that the far infrared pure rotation-translational spectrum be calculated for the temperatures and densities appropriate to the Earth's upper atmosphere. These results can then be included as empirical corrections to the absorption coefficient arising from allowed molecular transitions in a manner analogous to that used for the water vapor continuum or for aerosols[10].

Furthermore, the analysis should be extended to include the fundamental vibration-rotational band centered around 2330 cm^{-1}. The high frequency wing of this latter band (beyond 2400 cm^{-1}) is important even for paths in the lower atmosphere[3]. Since there is experimental evidence that the absorption in this spectral region depends on the absolute humidity, it will also be necessary to generalize the theory

to include the effects of dipole-induced-dipoles arising from N_2-H_2O collisions[11]. The final goal would be to calculate the collision-induced absorption for a realistic mixture of atmospheric gases.

Table I. Induction mechanisms used in the analysis of the far infrared spectrum of N_2

Induced Dipole (Radial Part)	Selection Rules	Transitions Included in the Analysis	Width Parameter δ at 300K in cm^{-1}
$\frac{\sqrt{3}\,\alpha Q}{R^4} + \lambda_3 e^{-R/\rho_3}$	$\Delta J=0,\pm 2$	Q(J),S(J),O(J)	15
$\frac{\gamma Q}{R^4}$	$\Delta J=0,\pm 2$	Q(J),S(J),O(J) S(J)+S(J') Q(J)+O(J') S(J)+O(J')	15
$\frac{\alpha\Phi}{R^6}$	$\Delta J=0,\pm 2,\pm 4$	Q(J),S(J),O(J) U(J),M(J)	22
$\frac{\gamma\Phi}{R^6}$	$\Delta J=0,\pm 2,\pm 4$	U(J)+S(J')	22
$\lambda_1 e^{-R/\rho_1}$	$\Delta J=0,\pm 2$	Q(J),S(J),O(J)	50

Table II. Dipole parameters used for calculating the far infrared absorption spectrum of N_2

Parameter	Value (atomic units)
Q	$-1.09\ ea_o^2$
α	$11.92\ a_o^3$
γ	$4.76\ a_o^3$
ϕ	$-10.4\ ea_o^4$
λ_3	$1.0 \times 10^{-3}\ ea_o$
$\lvert\lambda_1\rvert$	$1.0 \times 10^{-3}\ ea_o$
ρ_3	$0.765\ a_o$
ρ_1	$0.765\ a_o$

Table III. The natural logarithm of the spectral density function, $\ln G(\nu)$, at 200K for 10 cm^{-1} intervals from 0 to 600 cm^{-1}.

-15.25997, -15.40952, -15.66501, -15.71601, -15.70209, -15.69748,
-15.71930, -15.80301, -15.92511, -16.09787, -16.32028, -16.57281,
-16.87186, -17.19594, -17.54028, -17.91943, -18.29782, -18.67004,
-19.05135, -19.39171, -19.67974, -19.96159, -20.19299, -20.35023,
-20.57169, -20.77957, -20.91987, -21.14546, -21.37668, -21.56020,
-21.78352, -22.04158, -22.27176, -22.50496, -22.79155, -23.06866,
-23.35567, -23.73718, -24.13599, -24.51112, -24.86736, -25.21317,
-25.55823, -25.90428, -26.25133, -26.59933, -26.94826, -27.29811,
-27.64882, -28.00037, -28.35272, -28.70586, -29.05974, -29.41434,
-29.76962, -30.12556, -30.48212, -30.83928, -31.19699, -31.55525,
-31.91402/

REFERENCES

1. R.M. Goody, Atmospheric Radiation, (Oxford University Press, London, 1964); R.A. McClatchey, W.S. Benedict, S.A. Clough, D.E. Burch, R.F. Calfee, K. Fox, L.S. Rothman, and J.S. Garing, "AFCRL Atmospheric Absorption Line Parameters Compilation", AFCRL-TR-73-0096, 1973.

2. S.A. Clough, F.X. Kneizys, R. Davies, R. Gamache, and R. Tipping, "Theoretical Line Shape for H_2O Vapor; Application to the Continuum", in Atmospheric Water Vapor, (Academic Press, New York, 1980).

3. L.S. Bernstein, D.C. Robertson, J.A. Conant, and B.P. Sandford, "Measured and Predicted Atmospheric Transmission in the 4.0-5.3μm Region, and the Contribution of Continuum Absorption by CO_2 and N_2", Applied Optics, Vol. 18, pp. 2454-2461, 1979.

4. M.M. Shapiro and H.P. Gush, "The Collision-Induced Fundamental and First Overtone Bands of Oxygen and Nitrogen", Can. J. Phys., Vol. 44, pp. 949-963, 1966.

5. D.R. Bosomworth and H.P. Gush, "Collision-Induced Absorption in Compressed Gases in the Far Infrared II", Can. J. Phys. Vol. 43, 751-769, 1965.

6. J.D. Poll and J.H. Hunt, "Analysis of the Far Infrared Spectrum of Gaseous N_2", Can. J. Phys., Vol. 59, pp. 1448-1458, 1981.

7. J.D. Poll and J.L. Hunt, "On the Moments of the Pressure-Induced Spectra of Gases", Can. J. Phys., Vol. 54, pp. 461-470, 1976.

8. H.B. Levine and G. Birnbaum, "Classical Theory of Collision-Induced Absorption in Rare-Gas Mixtures", Phys. Rev., Vol. 154, pp. 86-92, 1967.

9. G.D. Billing and E.R. Fisher, "VV and VT Rate Coefficients in N_2 by a Quantum - Classical Model", Chem. Phys., Vol. 43, pp. 395-401 1979.

10. F.X. Kneizys, E.P. Shettle, W.O. Gallery, J.H. Chetwynd, Jr., L.W. Abreu, J.E.A. Selby, R.W. Fenn, and R.A. McClatchey, "Atmospheric Transmittance/Radiance: Computer Code LOWTRAN 5", AFGL-TR-80-0067, 1980.

11. J.D. Poll and R.H. Tipping, "Intensities of Collision-Induced Spectra of Diatomic Gases: Multipolar Single and Double Transitions", Can. J. Phys., to be published.

1982 USAF - SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

PHOTOIONIZATION OF IODINE MOLECULES AND CLUSTERS IN A SUPERSONIC MOLECULAR BEAM

Prepared by:	Dr. Edward A. Walters and J. Kent Newman
Academic Rank:	Associate Professor
Department and University	Department of Chemistry University of New Mexico
Research Location:	Air Force Weapons Laboratory, ARAP, Kirtland AFB, NM 87115
UASF Research	Capt. David Ellis
Date:	September 10, 1982
Contract No.	F49620-80-C-0035

PHOTOIONIZATION OF IODINE MOLECULES AND CLUSTERS IN A SUPERSONIC MOLECULAR BEAM

by

Edward A. Walters

and

J. Kent Newman

ABSTRACT

In preparation for a study of the photoionization of I_2, I_3, I_4, etc. a photoionization mass spectrometer employing a supersonic molecular beam emerging from a small nozzle was relocated from Los Alamos National Laboratory. Much of the effort described here has to do with moving and critical alignment of the mass spectrometer. Also described are the results of a literature survey of the photoionization and photoelectron spectroscopy literature on the important iodine molecules and clusters. Design of a nozzle for the generation of I_3 is discussed. Work done in interfacing a PDP 11/23 computer to the mass spectrometer and initial results of software development for controlling the experiments are described.

ACKNOWLEDGEMENT

The authors are pleased to extend their thanks to the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Research for providing them with the opportunity to spend 10 weeks this summer closely affiliated with the personnel and programs of the Air Force Weapons Laboratory, Kirtland AFB, NM. We are particularly grateful to the personnel in ARAP branch for their assistance and interest in this project. We especially wish to recognize Lt. Col. William E. McDermott for interesting and fruitful discussions leading to the identification and Captain David E. Ellis for continued and faithful support during this effort. We also acknowledge helpful discussions with Dr. R. Shea and assistance in locating a variety of pieces of equipment and hardware from Mr. R. Rathge.

I. INTRODUCTION

The iodine laser depends upon the production of electronically excited iodine atoms. Any channel for chemical reaction or energy transfer for which the velocity exceeds the stimulation of emitted radiation will reduce the laser efficiency. It is possible that the reaction

$$I + I_2 \longrightarrow I_3$$

is such a chemical reaction since I_3 has been calculated to be bound. If I_3 survives for more than even a few collisions, the laser efficiency will be affected. Thus, it is important to know and understand the properties of I_3. We report here the initial efforts in the direction of preparing and studying I_3 by photoionization mass spectrometry.

II. THE IODINE LASER

The means of generating excited iodine, I*, is a crucial step. There are two methods that have been examined rather extensively: flash photolysis of organic iodides (RI), and energy transfer from singlet oxygen ($O_2^1\Delta$). Both of these methods will be discussed briefly.

A. Flash Photolysis.

The iodine photochemical laser operates by flashlamp photolysis of an organic compound to produce iodine atoms in the upper state of the ground state multiplet, $I(5^2P_{1/2})$. The effective iodide has been perfluoroisopropyl iodide, $i\text{-}C_3F_7I$, though others may be used. A flashlamp emitting light in the range of 35,000 to 45,000 cm^{-1} excites $i\text{-}C_3F_7I$ to a dissociative upper level which leads almost exclusively to $I(5^2P_{1/2})$.

$$i\text{-}C_3F_7I + h\nu \longrightarrow i\text{-}C_3F_7 + I^*$$

The net result is the conversion of the absorbed uv photon with an energy of ~4.5 eV into an ir photon at an energy of ~1 eV in the process

$$I(5^2P_{1/2}) \longrightarrow I(5^2P_{3/2}) + h\nu(\lambda = 1.351\ \mu m)$$

The ground state iodine, $I(5^2P_{3/2})$, recombines very efficiently with $i\text{-}C_3F_7$ to regenerate the parent compound (1).

$$i\text{-}C_3F_7 + I(5^2P_{3/2}) \longrightarrow i\text{-}C_3F_7I.$$

There are a large number of secondary processes which affect the concentrations of atomic iodine and the radical C_3F_7. These include collisional deactivation of I*, radical recombinations of R•+R and R•+RI, radical-molecule reactions such as R•+RI and R•+I_2, and the three-body recombinations I*+I+M and I*+I+RI. Although many of the rate coefficients for these reactions are not known for R = C_3F_7, there are enough rate coefficients available for the closely related compound with R = CF_3 to permit reliable kinetic modeling. From a computed time development of concentrations based on these rate coefficients it has been shown that under typical operating conditions collisional deactivation of I* is not expected at times less than 10 μs.

Although the method of generating I* is different for the chemical iodine laser, the photophysics is nearly the same. Hence it is instructive to consider the levels involved in the laser in more detail.

The laser transition from $I(5^2P_{1/2})$ to $I(5^2P_{3/2})$ is a magnetic dipole transition for which the radiative lifetime is ~130 ms. For ^{127}I the degeneracy of both the upper and lower levels is removed by the magnetic dipole and the electric quadrupole moments of the nucleus which has a nuclear spin of 5/2. Consider first the upper level; the angular momentum can be parallel or opposed to the nuclear spin so the total angular momentum will be $F = 5/2 + 1/2 = 3$ or $F = 5/2 - 1/2 = 2$. For the lower level however there are four hyperfine states: $F = 5/2 + 3/2 = 4$, $F = 5/2 + 1/2 = 3$, $F = 5/2 - 1/2 = 2$, $F = 5/2 - 3/2 = 1$. The selection rule governing allowed transitions between these levels is $\Delta F = F_u - F_\ell = 0, \pm 1$. There are consequently six transitions allowed: $F_u - F_\ell = F_3 \rightarrow F_3$, $F_2 \rightarrow F_2$, $F_2 \rightarrow F_1$. The splitting of the upper levels F_2 and F_3 is larger than the splitting in the lower levels, so the six transitions cluster into two sets of three closely spaced transitions. The transitions center around 7603.15 cm^{-1} with a total frequency spread of 0.7 $cm^{-1} \approx 2 \times 10^{10} s^{-1} = 20$ GHz.

The Einstein coefficients A for stimulated emission have been reported for the six laser lines (2). These values are reproduced in Table I along with

Table I. Einstein coefficients and line center cross sections of the iodine laser lines.

$F_u \rightarrow F_\ell$	3→4	3→3	3→2	2→3	2→2	2→1
$A_{u\rightarrow\ell}, (s^{-1})$	5.0	2.1	0.6	2.4	3.0	2.3
$\sigma_{u\rightarrow\ell}(\nu_o)$, $(10^{-18} cm^2)$	6.0	2.4	0.66	2.67	3.3	2.55

the line-center cross sections, $\sigma_{u\to\ell}(\nu_0)$, for photolysis of i-C_3F_7I at 20 Torr. The net spontaneous emission rate for the process

$$I^* \longrightarrow I + h\nu$$

is given by

$$\text{rate} = A[I^*]$$

Here A is the effective Einstein A coefficient which is obtained by averaging individual A values over the initial (upper) states and summing over the final (lower) levels. The average over initial states is done by computing the degeneracies of the upper levels. Since degeneracy is 2F + 1 we get for $F_u = 3$ and $F_u = 2$ that $g_3 = 7$ and $g_2 = 5$, respectively. Thus the statistical weights are 7/12 for $F_u = 3$ and 5/12 for $F_u = 2$. These considerations give

$$A = \left(\frac{g_3}{g_2+g_3}\right)\left(A_{3\to4} + A_{3\to3} + A_{3\to2}\right)$$

$$+ \left(\frac{g_2}{g_2+g_3}\right)\left(A_{2\to3} + A_{2\to2} + A_{2\to1}\right)$$

Using the degeneracies calculated above and the $A_{u\to\ell}$ values from Table I, we get $A = 7.7\ s^{-1}$.

A parallel argument can be used to derive an expression for a net stimulated emission cross section $\sigma(\nu)$ as a function of frequency:

$$\sigma(\nu) = \left(\frac{g_3}{g_2+g_3}\right)[\sigma_{3\to4}(\nu) + \sigma_{3\to3}(\nu) + \sigma_{3\to2}(\nu)]$$

$$+ \left(\frac{g_2}{g_2+g_3}\right)[\sigma_{2\to3}(\nu) + \sigma_{2\to2}(\nu) + \sigma_{2\to1}(\nu)].$$

This is a very interesting result since the frequency dependence of $\sigma(\nu)$ and therefore the gain depends upon the extent of overlap and linewidths of the individual transitions. The linewidths can be manipulated by changing pressure and gas mixture in the laser cavity. This ability to control the frequency and intensity of stimulated emission is a remarkable and powerful feature of the iodine laser.

B. The Oxygen-Iodine Laser.

Excited iodine atoms can also be produced by energy transfer from molecular oxygen in the singlet state, $O_2(^1\Delta)$. There are two common ways of generating $O_2(^1\Delta)$. First, excited oxygen is produced in usable concentrations when a mixture of O_2 in He is passed through a microwave discharge at 2450 MHz. Alternatively, it is the product of the reaction between Cl_2 and a mixture of NaOH and H_2O_2. This later route provides a chemical pathway to pumping the iodine laser and consequently presents many attractive possibilities.

The lasers that have been constructed to date have been almost exclusively small-scale devices in which I_2 is injected into a cavity containing $O_2(^1\Delta)$. The net result of this interaction is the generation of I*, but the mechanism by which this occurs is unknown. It is generally believed that the reaction consists of at least two steps, the dissociation of I_2 by $O_2(^1\Delta)$ to give I*. This however is not at all clear. Nevertheless,

if the laser is to be scaled from a "bench-top" model to something larger, the order and mechanism of the reactions must be understood to assure proper scaling.

The first reported work on I_2 dissociation by $O_2(^1\Delta)$ was by Derwent, Kearns, and Thrush (3). In addition to the evidence of emission from atomic iodine they noticed a yellow glow which was attributed to the I_2 transition $B^3\pi_{0^+u} \rightarrow X^1\Sigma$. The mechanism they suggested to be compatible with their results was

$$O_2(^1\Sigma) + I_2(X^1\Sigma) \rightarrow O_2(^3\Sigma) + 2I$$

$$O_2(^1\Sigma) + I_2(X^1\Sigma) \rightarrow O_2(^3\Sigma) + I_2(A^3\pi_{1u})$$

$$I_2(A^3\pi_{1u}) + O_2(^1\Delta) \rightarrow O_2(^3\Sigma) + I_2(B^3\pi_{0^+u})$$

The serious flaw with this mechanism is that I* atoms are produced in the chemically pumped reaction and that under these conditions sufficient H_2O is present to quench all $O_2(^1\Sigma)$ before reaction with I_2 is possible. This was confirmed by experiments in which I* was produced in the absence of $O_3(^1\Sigma)$ (4). Thus the only other excited species, $O_2(^1\Delta)$, must be responsible for the dissociation and energetics requires at least two of these molecules. A number of possible mechanisms are currently under investigation, but they are all still quite speculative and very little information is available which even favors such a mechanism as, say, one involving vibrationally excited $O_2(^1\Delta)$ or the theoretically predicted but experimentally unknown dark state of $I_2(A'^3\pi_{2u})$.

In order to minimize the quenching effects of all the compounds found in the chemically pumped laser and thereby produce efficient cw operation of the oxygen-iodine laser, it is important to remove the products as quickly as possible from the active volume. This is accomplished by rapid mixing of the reactants $O_2(^1\Delta)$ and I_2. In order to obtain high cw power levels it is necessary that the gas flow velocities be supersonic. In the case of the HF/DF laser this is done by injecting H_2 or D_2 through nozzles into an expanding flow of F atoms in an appropriate carrier gas. Unlike the HF/DF system, for which a large amount of practical data are at hand, the problems of mixing a very massive molecule like I_2 into a relatively light gas of O_2 molecules have not been resolved. The low cross section for stimulated emission also dictates that the mole fraction of iodine be high. It is evident therefore that supersonic mixing introduces a number of other unanswered questions into the operation of the oxygen-iodine laser.

There are more questions related to the basic chemistry of the laser. For example, in a clean system of I atoms and $O_2(^1\Delta)$ molecules only the four following reactions are needed to describe the observed kinetics (5);

$$O_2(^1\Delta) + I \quad <==> \quad O_2(^3\Sigma) + I^*$$

$$I^* + O_2(^1\Delta) \dashrightarrow I + O_2(^1\Sigma)$$

$$I^* + O_2(^3\Sigma) \dashrightarrow I + O_2(^3\Sigma)$$

$$O_2(^1\Delta) + O_2(^1\Delta) \dashrightarrow O_2(^3\Sigma) + O_2(^1\Sigma)$$

Although there does exist data on all of these reactions, it is relatively

incomplete and, in fact, there is considerable uncertainty about the $I^* + O_2(^3\Sigma)$ reaction.

In a real system there will be contaminants in the flow, principally H_2O and Cl_2; rates of reactions of these species with I* and $O_2(^1\Delta)$ are not well known .

Iodine recombination rates are also critical (5). The reaction rates are known for the three-body process

$$I + I + I_2 \longrightarrow 2I_2$$

at room temperature to 1000°K and atmospheric pressure. Extrapolation to temperatures of 100-200 K, as encountered in supersonic nozzle expansions, is very uncertain except for the knowledge that the three-body rate increases as the temperature decreases.

Finally, cluster formation is a well-known phenomenon in supersonic nozzle expansions. The possibility of the formation of small metastable clusters in reactions like

$$I + I_2 \longrightarrow I_3$$

need to be explored. It has been predicted that I_3 is bound (6); if so, the reaction resulting in its formation will be a channel for depletion of the I atoms needed for producing high energy laser emission. This is but one of several broader questions related to cluster formation and nucleation phenomena in a supersonic oxygen-iodine laser (7).

a) How readily does nucleation occur?

b) What is the time scale for nucleation and what implications does this have for the O_2I^* laser?

c) Will wall or nozzle conditions influence nucleation?

D. Purpose of this Work.

In light of the preceding discussion, it was agreed that a better understanding of I_3 is important. This species has not been observed experimentally, so a somewhat exotic approach is needed. We decided to try to generate I_3 in a supersonic nozzle source and to study it by photoionization mass spectrometry. A photoionization mass spectrometer with a supersonic nozzle was available for loan from Los Alamos National Laboratory, so the goal of this effort was (a) to move the apparatus from Los Alamos National Laboratory, (b) to design nozzles for the generation of I_2 and I_3, and (c) to automate the data collection and reduction steps by interfacing the apparatus with a PDP 11/23 computer. We report the results of the effort.

III. I_2 PHOTOIONIZATION

The absorption spectrum of the iodine molecule has been studied extensively over a period of many years. The great majority of this work has concentrated on the region above 1700 Å and surprisingly little has been done in the vacuum-UV region below this value. There are several reasons for this. First, autoionization features are dominant from threshold into the far UV thereby making identification of vibrational progressions and ionization to excited states difficult. In addition, the close spacing of the vibrational levels of I_2, 214 cm^{-1}, means that at 300 K there is considerable population of the v=2 and v=3 levels, so hot bands obscure the ionization threshold and other vibrational sequences. The ionization potential has been

determined by vuv spectroscopy (8), electron impact ionization (9), photoelectron spectroscopy (10-12), photoionization without mass selection (13-15) and photoionization with mass analysis (15,16). Photoion yield curves for the region of 1050 to 1410 Å (13,16) have been published. In the first instance (13) only total ion current was measured, so the rather prominent structures at lower energy than the spectroscopic ionization potential of 9.400 eV (8) were attributed to the ion-pair formation

$$I_2 + h\nu \longrightarrow I_2^* \longrightarrow I^+ + I^-.$$

The threshold for this process was measured as 8.84 eV. This corresponds well with the earlier results of Morrison et al. (15) and Watanabe (17) who give 8.85 ± 0.1 eV and 8.93 ± 9.02 eV, respectively, as the onsets for ion-pair production. Ionization potentials for I_2 have been established from these reports as being close to 9.38 eV.

The photoion yield curves are characterized by extensive autoionization features which largely obscure ionization to the higher electronic states of I_2^+ that have been identified by PES. In addition, spin-orbit splittings and vibrational transitions cannot be clearly identified. Because of the relatively low vapor pressure of I_2 (0.35 Torr at 300 K), no effort was made to examine the photoion yield curve close to threshold as a function of temperature for the purpose of clarifying which of the observed peaks are due to hot bands.

On the basis of our experience with supersonic nozzles, however, it is reasonable to expect the cooling of the gases during expansion to be sufficient to reduce or even eliminate the problem of hot bands (18-20). With the resolution of our apparatus, 0.006 meV for monomers, and the spectral

range accessible, 960-1250 Å, a detailed exploration of the photoion yield behavior of I_2 is worthwhile. The autoionization behavior should be resolvable into Rydberg series that can be identified through convergence to the second and higher ionization limits. The photoion yield curve for the ion pair production partner I^+ should also provide very interesting results for detailed analysis. These thoughts are the incentive for some of the photoionization studies we have initiated this summer. A thorough survey of the literature revealed nothing pertaining to the cluster species $I \cdot I_2 = I_3$. Consequently, any information that can be supplied from photoionization will be a useful contribution.

IV. APPARATUS RELOCATION AND RECONSTRUCTION

The primary objective of the summer work was to relocate a molecular beam photoionization mass spectrometer from Los Alamos National Laboratory to a permanent location at the University of New Mexico. In preparation for the move, a large number of detailed photographs of the assembled apparatus were taken to assure ease of reassembly and relocation of wiring, flanges, fittings, forepumps, diffusion pumps, etc. These were printed as 8" x 10" glossies which have subsequently been used extensively. Figures 1-3 are photographs of the assembled apparatus at LANL. Disconnection of the electrical components and labeling of the wires were completed in early June. At this time all parts of the apparatus were checked for possible radioactive contamination by H Division personnel. Direct counting was done where possible, smears were taken of less accessible portions, and water lines were flushed with dilute acids which were then monitored. No evidence of any residual radioactive material was found. Next, the mass spectrometer was carefully disassembled and all critical parts -- nozzle source, ionizer, grating, channeltron -- were

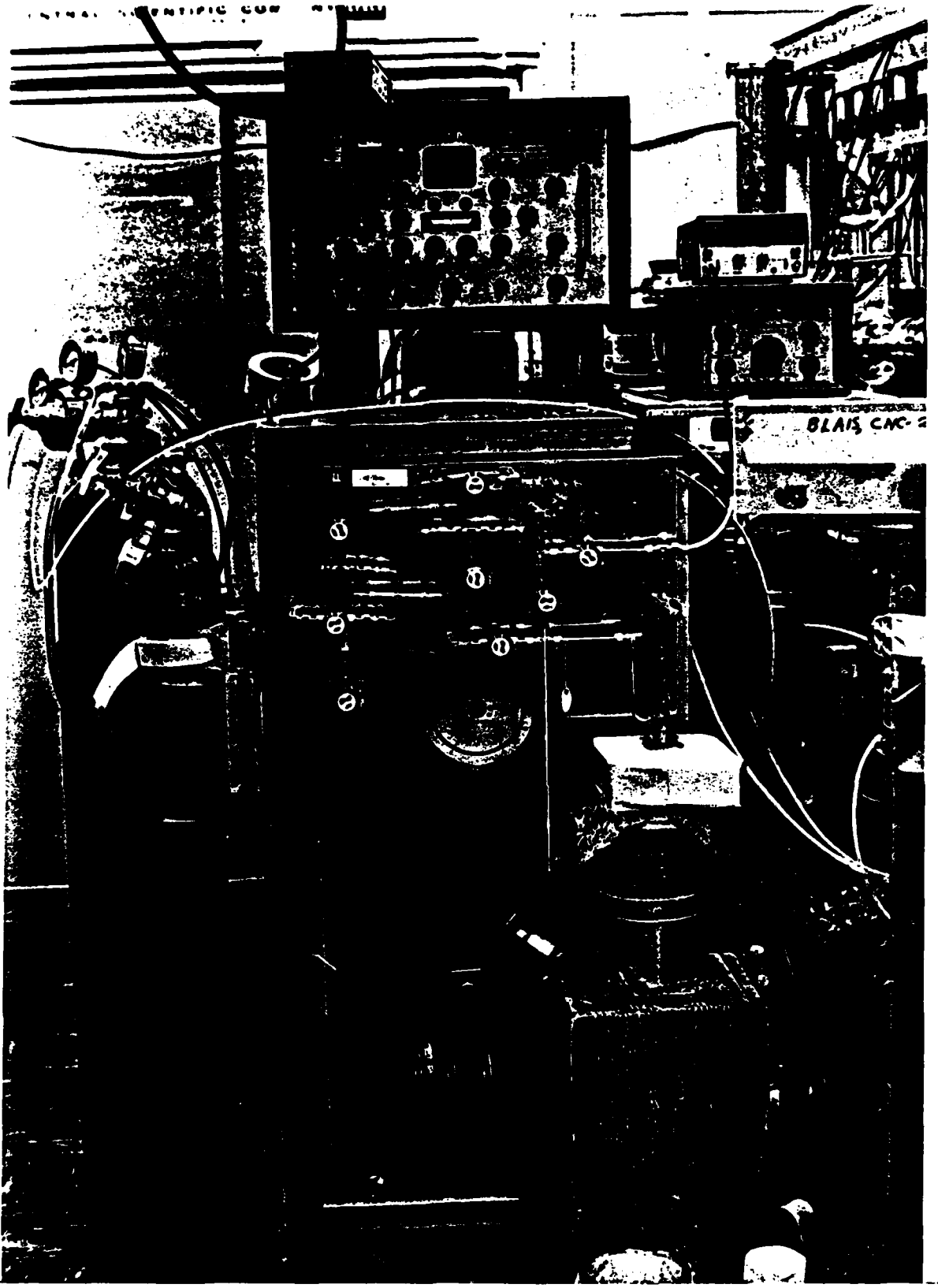


Figure 1. Gas inlet manifold and magnetic field strength monitoring apparatus on photoionization mass spectrometer.

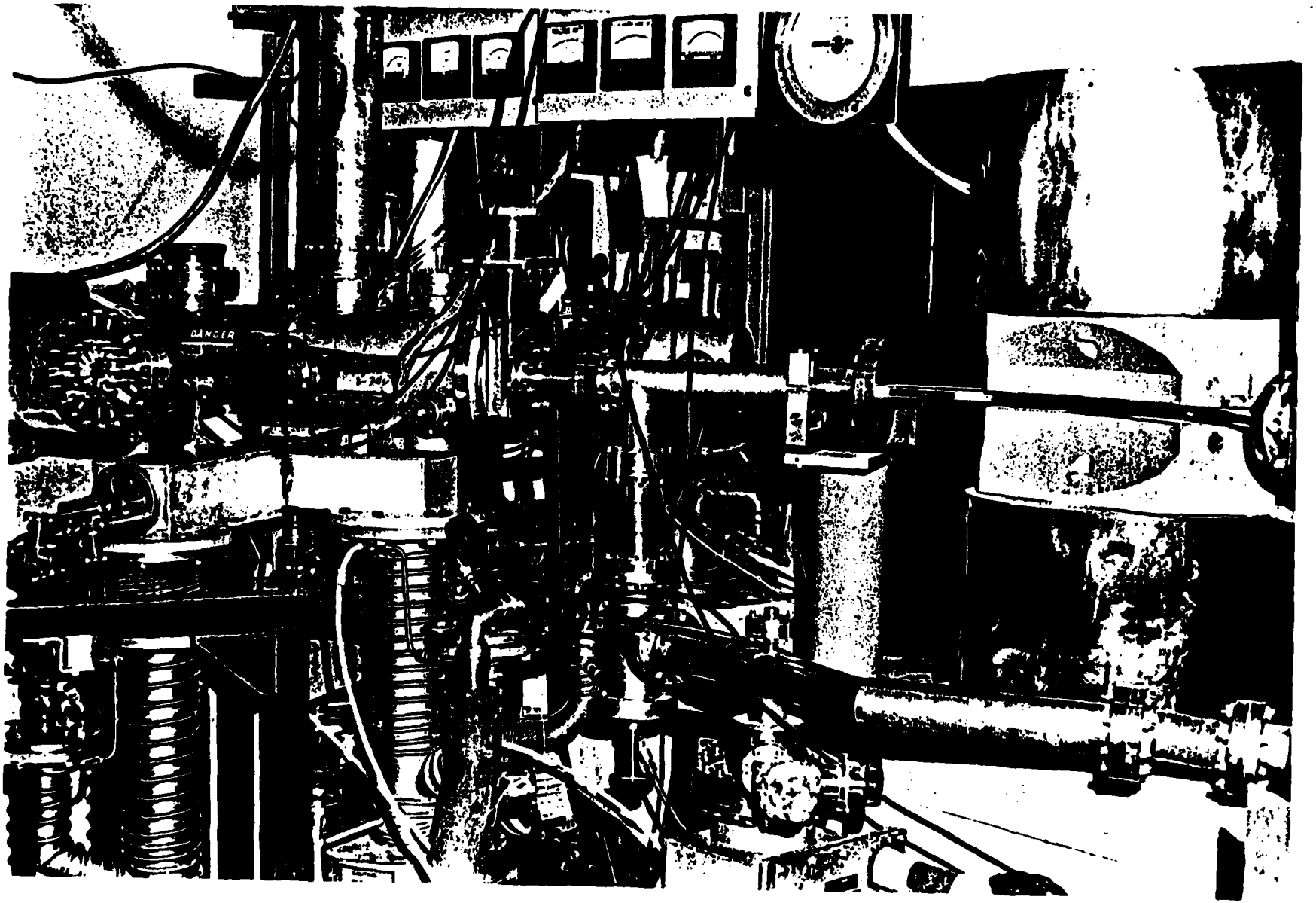


Figure 2. Source region of the photoionization mass spectrometer. The electromagnet appears on the right hand side of the photograph.

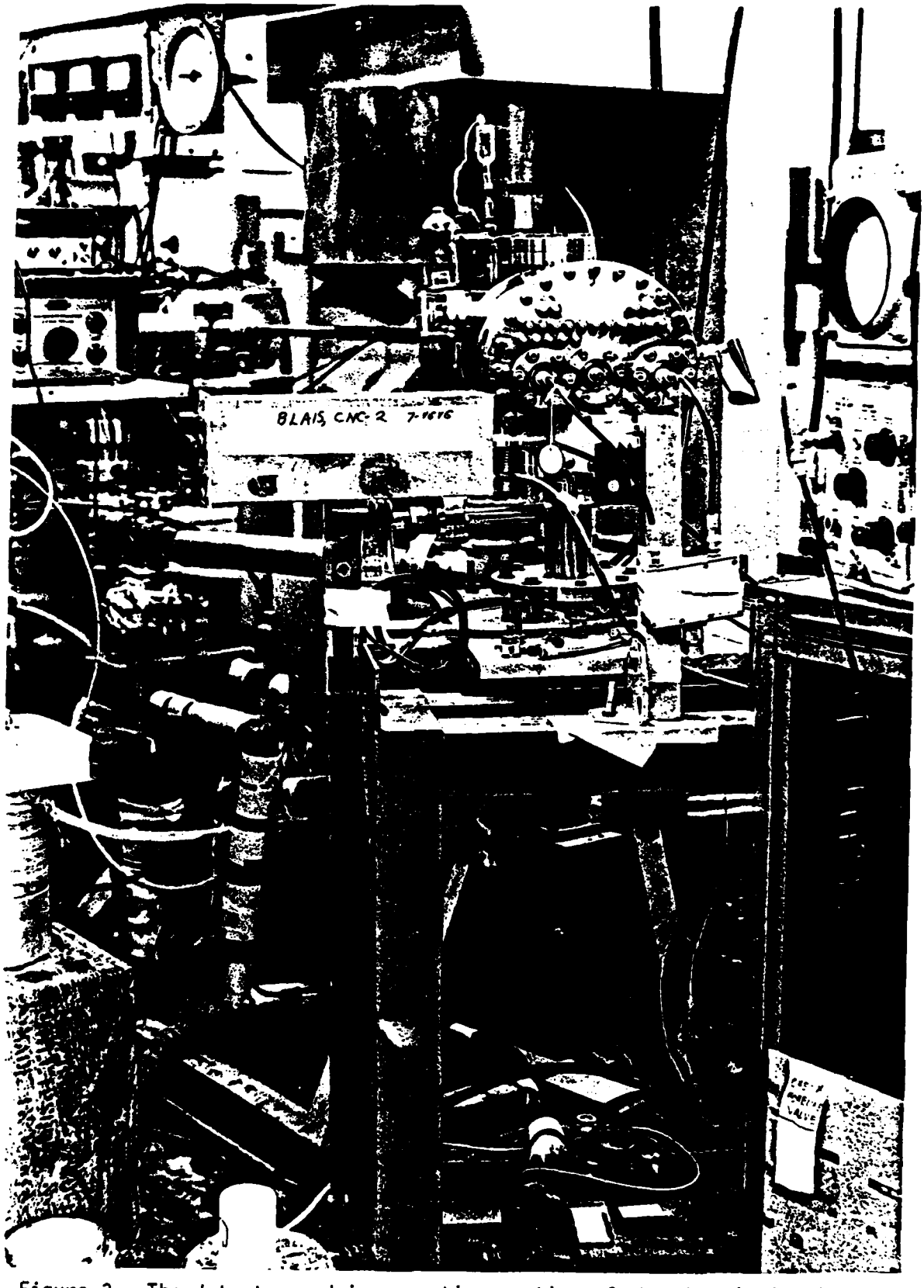


Figure 3. The detector and ion counting portion of the photoionization mass spectrometer.

packed with extreme care. This was then moved to UNM by U-Haul truck. This was accomplished by early July.

The heart of the apparatus is the electromagnet which weighs about 6,000 lbs. Special provisions and coordination of effort was required for moving the magnet. For removing the magnet and stand from the CMR Building at LANL, Zia Corporation provided a crew of riggers; they placed the magnet on a truck from UNM which was used to transport it to Albuquerque. The UNM Physical Plant service lifted the magnet from the truck and set it on the ground next to the Chemistry Building. The magnet is too heavy for the freight elevator in the building so Crane Services, Inc. was hired to move the magnet into the lab and mount it on the stand. Because of the awkward location of the balance point on the magnet, it was necessary to fix a set of I-beams to the top of the magnet for lowering it into a doorway with direct access to the basement. Once in the lab another arrangement of I-beams was needed to lift the magnet into position so the stand could be moved under it. This was accomplished by using a set of short (8") hydraulic jacks mounted on cribbing of large timbers. This last stage of getting the magnet into the building and mounted properly took about 1 week of full-time effort.

While this was being done the source chamber was reassembled, leak-checked, and pumped down. With very little effort besides closing all the flanges, the source chamber was pumped down to a pressure of 5×10^{-7} Torr. This assures us that the apparatus is tight enough to begin experiments, that is, it suffered no major damage in the move. A beam of He atoms was run through the nozzle and routine checks were performed on the beam flag, ion gauges, power supplies, etc. which showed that all are in working order.

We also performed some checks with the monochromator. Since the monochromator as originally designed does not fit into the available space around the ion source compartment, it was necessary to reverse the entrance and exit arms. This means that the grating must be turned so the blazed grooves point in the correct direction. It also means however that to scan into the +1 order, the scan must be in the opposite direction from the zero-order reflection than the readout drive on the monochromator. Some simple experiments confirmed this expectation. Slight modification of the monochromator will be required when it is being mounted to assure maximum recovery of light from the H_2 discharge lamp.

Magnet Alignment Considerations. The first, and extremely critical, job is proper alignment of the source and detector chambers. This is essential if adequate resolution is to be achieved. The resolution R is given by

$$R = \frac{r}{2(s_1+\psi(r))}$$

where r is the radius of curvature of the magnet (17.750") in this case) and s_1 is the source exit slit width. $\psi(r)$ is the error term which can be 0.004", expressed as

$$\psi(r) = r\alpha^2 + E$$

where α is the half angle of the lateral divergence out of the source slit. In this apparatus α is determined by the quadrupole focusing lens system and is 0.015 radians. E is the sum total of errors which arise from misorientation of the magnet, anomalies in the magnet field, etc. For our

apparatus $r_\alpha^2 = 0.004"$, so that if a resolution R = 1000 is desired, E cannot exceed 0.0018". We are confident that this maximum value for E has been achieved with this mass spectrometer since the mass spectrometer equation

$$m/e = \frac{B^2 r^2}{2V}$$

has been previously satisfied. In this equation m is the mass of the particle, e is the electrical charge, B is the magnetic field strength, V is the accelerating potential, and r is the particle orbit radius. Figure 4 shows the relationships of the slits and the magnet with respect to one another. The desired resolution (R = 1000) will be obtained when the following conditions are satisfied:

a) The distance between the slits if 71.000±0.005".

b) $X_1 = X_2 \pm 0.005"$.

c) The ions enter the magnetic field at an angle of 90° ±0.001 rad.

d) The mass spectrometer equation is satisfied.

The cross-hatched area in Figure 4 indicates the location of the virtual magnet. By rule-of-thumb, the outer edge of the fringing field is one pole gap away from the side of tne magnet. The pole gap in our magnet is 7/8". Although the radius of curvature of the real magnet is 16", when the fringing field is taken into account the radius of curvature of the virtual magnet is r = 17.750". The geometry of the system then requires that $X_1 = X_2 = r\sqrt{3}$ + 0.875" = 31.618". This condition must be met in the first-order alignment, second-order alignment will later be done by moving the magnet just enough so that mass spectrometer equation is satisfied.

Alignment Procedure. In order to accomplish the alignment several special pieces of equipment were made. These were:

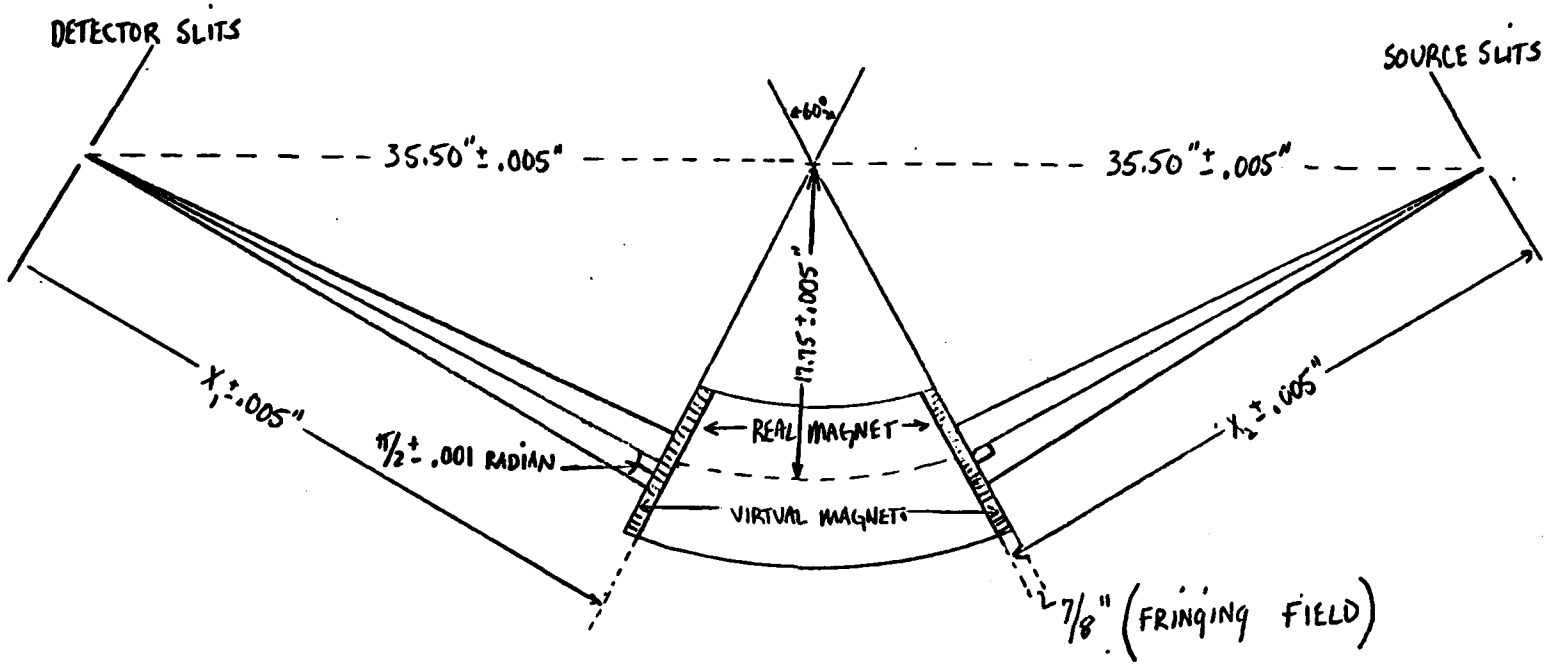


Figure 4. A schematic diagram at the magnet and the required orientations with respect to the source and detector slits.

a) An aluminum bar whose total length is 71.000±0.010". The center is clearly marked and the ends are oriented at an angle so the tips fit into the slits of the source and detector.

b) Two pieces of aluminum angle were fit with threaded rods to hold them against the sides of the magnet. The center of curvature of the virtual magnet was precisely located near the intersection of these bars.

c) A block of aluminum was machined to fit between the pole faces of the magnet. A long rod was calibrated so the distance between the magnet sides and the slits could be determined to ±0.005".

d) A second aluminum block was manufactured to fit between the magnet pole faces. This one was cut so the side facing the slit is perpendicular to the lower magnet pole face to ±0.001 rad. A first surface mirror was attached to this face.

e) A surveyor's transit.

These tools were then used to complete the alignment. First, the center of curvature of the virtual magnet was located near the intersection of the two aluminum angles mounted on the magnet sides. Second, the long aluminum bar was centered over this point so that it was perpendicular to the symmetry axis of the magnet. The source and detector chambers were then moved until the tips of the bar touched the slits. Third, the distances from the magnet sides to the slits were measured and the chambers were adjusted to get these distances (X_1 and X_2) nearly identical. Fourth, the block with a mirror was placed between the poles faces of the magnet and viewed through the slits at both the source and detector ends with the surveyor's transit. These units were then oriented carefully and cautiously until they were symmetrically disposed about the line of sight through the telescope and the reflection of the slit was also symmetric about the line of sight. These steps were

repeated in this sequence until self-consistency of all the measurements was obtained. In the final configuration $X_1 = X_2 = 31.581 \pm 0.005$".

Reassembly. After the first-order alignment was obtained, the channeltron was fit into the detector chamber and entrance and exit slits were set at 0.004". The vacuum system of the analyzer was assembled and leak checking commenced. It appears that everything is in reasonably good order and that only a few O-rings and new pump oil will be needed before the desired pressure of $\sim 1 \times 10^{-7}$ Torr will be achieved.

V. I_3 NOZZLE DESIGN

The production of the possible intermediate I_3 is of critical importance to this experiment. The goal is to be able to generate a sufficient amount of I_3 in a supersonic molecular beam to be able to study it by photoionization. The most convenient way to do this would be to generate the I_3 in the nozzle itself, so the first attempt at constructing an I_3 nozzle is based on mixing I_2 and I immediately prior to expansion through the nozzle. The nozzle designed is being constructed according to the drawing given in Figure 5. It consists of a pair of concentric nozzles; I_2 will pass through the inner nozzle into a buffer region between the nozzles where the process $I + I \longrightarrow I_3$ can occur. The I_2 will be mixed with a carrier gas, probably helium or argon, and the entire nozzle and inlet system can be heated.

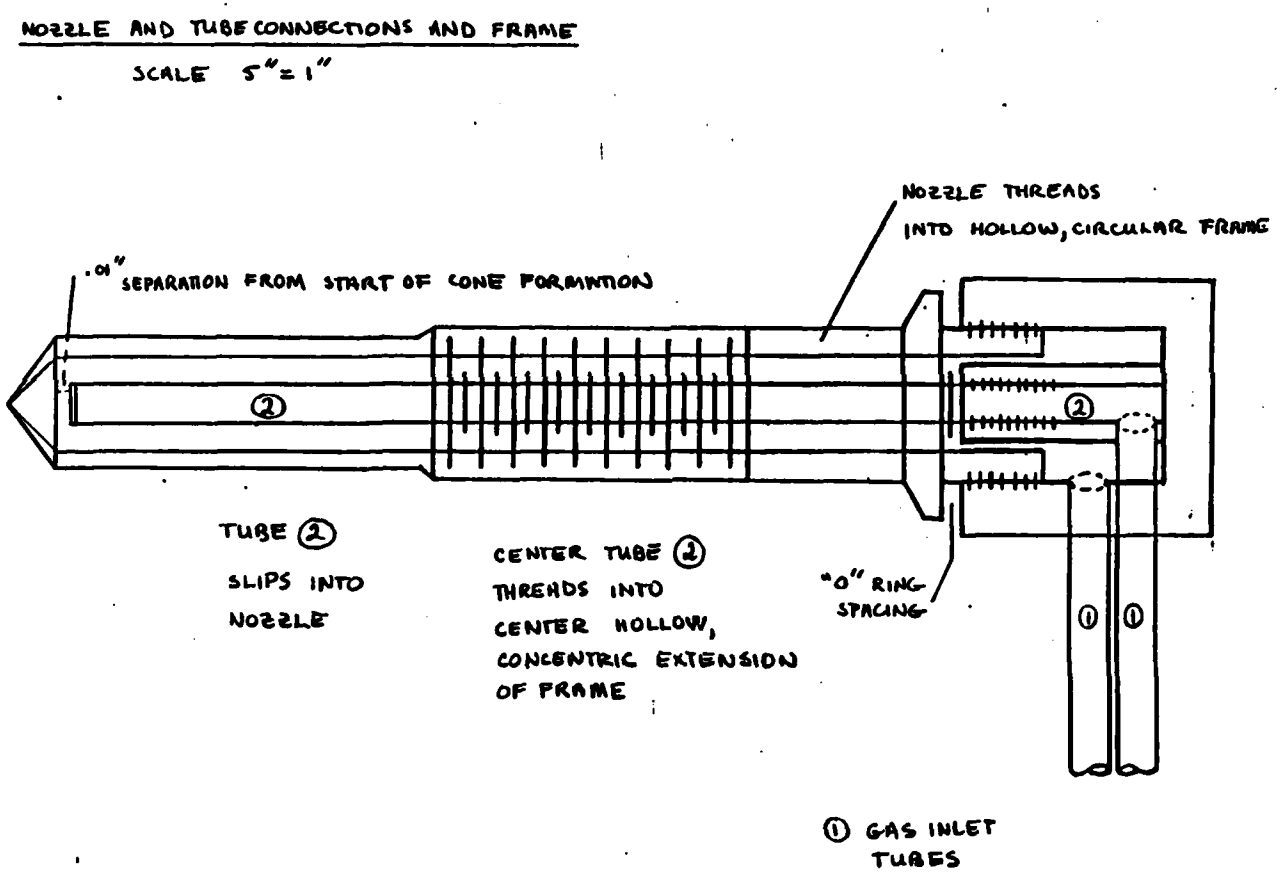


Figure 5. Schematic drawing of the nozzle for I_3 generation.

The outer portion is where I atoms are transported to the interaction region. The plan is to prepare I by the reaction

$$I_2 + O_2\ (^1\Delta) \longrightarrow 2I + O_2(^3\Sigma)$$

Singlet oxygen $O_2(^1\Delta)$ will be prepared by a microwave discharge through a flowing stream of O_2 in He. This will intersect the I_2 in an interaction region immediately prior to expansion. This nozzle design can be employed in a variety of ways to dissociate iodine. Much work will be required to characterize the nozzle and to establish optimum operating conditions. A preferred characterization technique is laser induced fluorescence.

If a satisfactory I atom concentration cannot be achieved by this approach, it may be necessary to revert to the standard nozzle design and to generate I atoms from photodissociation of I_2 or $(I_2)_2$. This possibility is suggested by the work of Valentini and Cross (21) who have reported substantial cage effects in I_2 Ar clusters when photodissociated at 488 nm with a cw argon ion laser.

VI. COMPUTER INTERFACING

The photoionization mass spectrometer was operating at Los Alamos National Laboratory required manual data collection and wavelength positioning in addition to manual data manipulation to correct for background, pressure drift, magnet drift, and normalization of day-to-day data. All of this is rather tedious and inefficient when the instrument is fully operational since the data reduction may require considerably more time than the data collection and the possibility of error from manual treatment of the data increases substantially. Consequently, we have proceeded with efforts to interface the equipment to a computer which will be used to do much of the routine work.

To this end we have acquired a PDP-11/23 computer with 32 K of memory. It uses the RT 11 operating system for data acquisition. The mass storage medium is a TU-58 cassette tape system. We expect to have access to either an RX01 or RX02 disc drive in the near future to use for software development, but these have not been available to us yet.

We have developed a parallel digital I/O to the photon counting system. This channel reads the total number of counts collected on a PAR Model 1112 digital synchronous computer in a preset period of time and resets the time to zero and initiates counting. This much of the code is currently working. We plan to include an option in which it is possible to choose between a predetermined counting time; a total number of counts such that the counting errors reach some selected percentage error given by $1/\sqrt{N}$, where N is the number of pulses collected; or the choice between $1/\sqrt{N}$ and some maximum count time if N is small. To write the software for these options, it is most useful to be able to connect the computer directly to the apparatus. We expect to be able to do this very shortly — we need only to find all the leaks in the analyzer section of the machine and then we can apply power to the detector and receive counts.

The computer has eight double-ended A to D channels which can be used for data collection. For example, it is important to monitor the intensity of light from the H_2 source as measured by the photomultiplier tube. The ideal situation is to, in fact, integrate the light over the counting period so that the number of ions produced per photon intersecting the molecular beam can be determined. This can be done easily with the computer. Also, it will be useful to monitor the various pressure gauges to be able to normalize the count rate to some pressure. This is important since the concentration of clusters in the beam is related to pressure by a function like $p_2 = p_T^n$ where n is some number greater than unity, p_T is the total pressure and p_2 is the partial pressure of dimers.

The computer system supports D-to-A channels which can be used to drive a stepping motor which will be attached to the monochromator drive system. This will allow us to advance the monochromator automatically by punching the desired wavelength increment into the computer console. This capability will also be used to display the final adjusted results on an XY plotter.

In preparation for writing our own software we have reviewed the code used on the photoionization apparatus at the National Synchrotron Light Source at Brookhaven National Laboratory, which was graciously loaned to us by J. R. Grover (22).

VII. RECOMMENDATIONS

It is quite evident that many uncertainties exist with respect to the operation of the oxygen-iodine laser. A large number of these questions pertain to the basic science of the reaction mechanism while a collection of others are related to the operation of the device itself. We have chosen to address some of the questions about nucleation or condensation phenomena in the oxygen-iodine laser. The experiments initiated are difficult because the species we wish to investigate, most notably I_3 and I_4, are short lived and transient. We have made considerable progress in constructing an apparatus which can supply information about these species, however in the 10 week period of this project we were not able to actually come up with answers. In view of the potential importance of these answers to the supersonic oxygen- iodine laser, we recommend that the photoionization and I_3 nozzle work be pursued. The limited experiments done on the photoionization of I_2 point to the need to clarify these results, particularly in supersonic beams. Clusters can be identified and examined by electron impact ionization is a rather crude way, and in much more detail with photoionization. Particular emphasis should be placed on obtaining thermodynamic parameters for the clusters and cluster ions. Development of an I_3 nozzle should be pursued.

REFERENCES

1. A. Ben-Shaul, Y. Haas, K. L. Kompa and R. D. Levine, "Lasers and Chemical Change", (Springer-Verlag, Berlin) 1981, 305-314.
2. K. Hohla and K. L. Kompa in "Handbook of Chemical Lasers", R.W.F. Gross and J. F. Bott (eds.) Wiley, New York) 1976, 667.
3. R. G. Derwent, D. R. Kearns and B. A. Thrush, Chem. Phys. Lett. 6, 115 (1970).
4. P. D. Whitefield and S. J. Davis, Laser Digest, AFWL-TR-80-4, 175 (1980).
5. R. F. Shea, personal communication.
6. H. Michaels, personal communication.
7. Lt. Col. W. McDermott, personal communication.
8. P. Venkateswarlu, Can. J. Phys. 48, 1055 (1970).
9. D. C. Frost and C. A. McDowell, Can. J. Chem. 38, 407 (1960).
10. D. C. Frost, C. A. McDowell and D. A. Vroom, J. Chem. Phys. 46, 4255 (1967).
11. A. B. Cornford, D. C. Frost, C. A. McDowell, J. L. Ragle and I. A. Stenhouse, J. Chem. Phys. 54, 2651 (1971).
12. S. Evans and A. F. Orchard, Inorg. Chem. Acta 5, 81 (1971).
13. K. Watanabe, T. Nakayama and J. Mottl, J. Quant. Spectrosc. Radiat. Transfer 2, 369 (1962).
14. J. A. Mayer and J.A.R. Samson, J. Chem. Phys. 52, 716 (1970).
15. J. D. Morrison, H. Hurzeler, M. G. Inghram and H. E. Stanton, J. Chem. Phys. 33, 821 (1960).
16. V. H. Dibeler, J. A. Walker, K. E. McCulloh and H. M. Rosenstock, Int. J. Mass Spectrom. Ion Phys. 7, 209 (1971).
17. K. Watanabe, J. Chem. Phys. 26, 542 (1957).

18. W. M. Trott, N. C. Blais and E. A. Walters, J. Chem. Phys. 69, 3150 (1978).

19. W. M. Trott, N. C. Blais and E. A. Walters, J. Chem. Phys. 71, 1692 (1979).

20. E. A. Walters and N. C. Blais, J. Chem. Phys. 75, 4208 (1981)

21. J. J. Valentini and J. B. Cross, J. Chem. Phys. 77, 572 (1982).

22. J. R. Grover, personal communication.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

DATA COMMUNICATION BETWEEN CDC CYBER 750 AND HP 1000

Prepared by:	Dr. Kai Wang
Academic Rank:	Associate Professor
Department and University	Department of Mathematics Wayne State University
Research Location:	Air Force Wright Aeronautical Laboratories Aero Propulsion Laboratory Turbine Engine Division Components Branch
USAF Research Colleague:	Dr. Kervyn Mach
Date:	August 7, 1982
Contract No:	F49620-82-C-0035

DATA COMMUNICATIONS BETWEEN CDC CYBER 750 AND HP 1000

by

Kai Wang

ABSTRACT

There are many different computers in use today. Each has its own relative strength and preferred peripherals. Establishing communications between two (or more) different computers can combine the best features of each machine to form a system with much more power and flexibility than a stand-alone computer can provide.

In our project, we have established data communication between CDC CYBER 750 and HP 1000 computers. We have developed a pair of programs which consists of a PASCAL program for the HP 1000 and a FORTRAN program for the CYBER 750. For downloading, the programs must be running concurrently to provide the necessary cooperation. Suggestions for further extending this system are offered.

ACKNOWLEDGEMENTS

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern center for Electrical Engineering Education for providing him with the opportunity to spend a very worthwhile and interesting summer at the Aero Propulsion Laboratory, Wright-Patterson AFB, Dayton, Ohio. He would like to acknowledge the Laboratory for his hospitality and excellent working conditions.

Finally, he would like to thank Dr. Kervyn Mach for suggesting this area of research and for his collaboration and guidance, and he would like to acknowledge help extended by Dr. Jim Petty.

I. INTRODUCTION: In our project, what we want to accomplish is to establish data communication between a CYBER 750 and a HP 1000 so that we can retrieve a file from the CYBER to be stored on the HP 1000's disk, and we can also send a file which was stored on the HP 1000's disk to the CYBER 750. Communication is to be initiated by a user of the HP 1000 whom the CYBER will perceive as just another user logged-on at a time-sharing terminal. The process of transmitting a file from the CYBER to the HP 1000 is termed 'downloading' and its converse is called 'uploading'. Since both the CYBER and HP 1000 are time-sharing systems with multiple users and a telephone-line link is used, it is impossible to avoid noise in the channel which make communication difficult. The task is also made more difficult because there are many differences (e.g. operating speed, input-output formats) between the CYBER 750 and HP 1000. It seems unlikely that one could obtain a data communication which is both fast and error-free. Thus, one has to sacrifice speed of communication and tolerate some possible errors. At present, we have established a data communication between the CYBER 750 and HP 1000 which is reasonably fast and develops relatively few errors.

For downloading, we have a pair of programs which consists of a PASCAL program for the HP 1000 and a FORTRAN program for the CYBER 750 which have to be running concurrently in order to establish cooperation. For uploading, a single PASCAL program for the HP 1000 is sufficient for the job. Both PASCAL programs are combined into a single program which also includes some other convenient features.

II. OBJECTIVES OF THE RESEARCH EFFORT: The main objective of this project was to establish data communication between a CYBER 750 and a HP 1000 so that they form a system. A disk file from the HP 1000 can be sent to the CYBER for storage in a specified disk file and a disk file from the CYBER can be retrieved and stored in a specified disk file on the HP 1000.

III. DESCRIPTION OF METHODS: The basic step in setting up the communication between the CYBER and HP 1000 is to interpret the message sent by the CYBER. The CYBER transmits characters in eight-bit bytes consisting of a parity bit followed by a seven-bit character code. The interface software in the HP 1000 stores two of these bytes in a 16-bit HP word. However, the HP 1000 assumes the leading (parity) bit to indicate a binary complement, so an algorithm is needed to convert each word into two ASCII characters. I am indebted to Dr. J. Petty for this algorithm.

The second step is to set up the initial contact with the CYBER by sending a "LOGIN" message and other necessary commands in order to attach the file to downloaded and load and run the FORTRAN program on the CYBER. The PASCAL program prompts the user for the identity of the file be downloaded, then automatically issues the necessary commands to the CYBER.

The FORTRAN program, on request from the PASCAL program on the HP 1000, provides the necessary information about the download file as the number of characters in each line and the check number of each

line (which is the total numerical value of the ASCII representation of the characters modulo 100) and sends the download file in blocks of 50 lines.

While sending each line, the CYBER adds at the beginning of each line a group of six control characters consisting of a carriage return, a line feed and four nulls. These must be removed before writing the line to the HP disc. The easiest way to do this is to pack the incoming data into a large string of characters and then check each character. In this way, we are able to handle up to 25 lines, but without any error checking facilities because of the limited memory space in the HP 1000.

To work around this limitation, we employ the following approach: We store each incoming line into an array of characters and strip off the control characters in the process. Before storing this line on the disk we can check the check-value to see if there are errors. If there are errors, then we request the CYBER to send the same block again and we ignore the lines which are already stored on the disk, down to the pending line. To avoid the danger of an infinite loop, we are content with ten tries to correct the errors for each block. If a line still has an error, it is indicated for later manual correction. Because of the resulting savings in memory, we are able to handle 50 lines each time.

IV. RECOMMENDATIONS: The program should be improved so that it can start downloading the file from the CYBER at any block or even any line because of the possible interruption while downloading a large data file.

1982 USAF-SCEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

ISOLATION OF PYROGENIC EXOTOXIN C FROM STAPHYLOCOCCUS AUREUS STRAINS ASSOCIATED WITH TOXIC-SHOCK SYNDROME

Prepared by:	Dr. Edward R. Ward, Jr.
Academic Rank:	Associate Professor
Department and University:	Department of Biology Alabama A&M University
Research Location:	USAF School of Aerospace Medicine Epidemiology Division Microbiology Branch
USAF Research Colleague:	Dr. Louis E. Blouse, Jr.
Date:	August 11, 1982
Contract No:	F49620-82-C-0035

ISOLATION OF PYROGENIC EXOTOXIN C FROM STAPHYLOCOCCUS AUREUS STRAINS ASSOCIATED WITH TOXIC-SHOCK SYNDROME

by

Edward R. Ward, Jr.

ABSTRACT

The development of a reproducible and clinically useful technique for the isolation and identification of pyrogenic exotoxin C (PEC) from isolates of Staphylococcus aureus is investigated. Results show that isoelectric focusing of extracellular proteins for one hour and subsequent staining of gels with silver stain to be the method most sensitive and reproducible. The methods used by the Centers for Disease Control (Dr. James Feeley, personal communication) and Schleivert et. al., (1981) were modified to increase sensitivity in detecting small amounts of PEC.

In addition, four isolates of S. aureus from different anatomic sites of the same individual with clinically diagnosed toxic-shock syndrome were examined with this technique. Of the four isolates (urine X2, throat, and cervix) only the cervical isolate showed PEC when compared to the positive controls.

Acknowledgement

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to spend an interesting and worthwhile summer at USAF School of Aerospace Medicine, Brooks AFB, TX. He would like to acknowledge the Epidemiology Division, in particular the Microbiology Branch for its hospitality and excellent working conditions.

Finally, he would like to thank Dr. Louis Blouse for suggesting this area of research and for his collaboration and guidance, and he would like to acknowledge the many helpful discussions with Mr. James Mitchell, Capt. Barbara Schramm, and the other members of the Microbiology Branch.

I. INTRODUCTION:

Toxic shock syndrome (TSS), a disease predominately of young women, has been associated with vaginal carriage of Staphylococcus aureus.[1,2,3] The clinical features of this acute illness are fever, marked hypotension, diarrhea, conjunctivitis, myalgias and a scarlatiniform rash followed by fine desquamation.[1,3,4] Blood cultures, throat cultures for group A streptocci, and viral cultures are usually negative.[5] Although tampon usage has been a major host characteristic in the recent epidemic of TSS, S. aureus infection appears to be associated with all well documented cases of TSS. In one study,[3] S. aureus was recovered from the vaginal fluid of all menses-related patients. There have been reports since 1927 of associations between clinical syndromes similar to TSS and S. aureus infections.[6,7,8] Also, TSS has been reported to occur in males with localized S. aureus infections.[1,2,5] The pathogenesis of TSS remains poorly understood, however. The typical serious systemic manifestations of this local infection suggest that toxin or toxins are involved. Schlievert et. al. (1981) were able to demonstrate that one of the exotoxins elaborated by TSS strains of S. aureus enhanced the effects of endotoxic shock in rabbits.

The possibility that isolates of S. aureus from several anatomic sites of a single individual may contribute to TSS, has not been investigated previously. Methods which could detect and identify characteristic extracellular protein markers (toxins) produced by S. aureus isolates cultured from patients with clinical TSS would be of great value to clinicians in establishing a definitive diagnosis.

The value of such methods to the Air Force Medical Service specifically is obvious.

The clinical signs and symptoms of TSS are not pathogonomic but, instead, may be seen in a variety of illnesses which may result in shock. In addition, recent epidemiologic investigations have shown that TSS is often a recurrent illness in women which may be prevented by anti-staphylococcal drugs and discontinued use of tampons.[2] Further, the full clinical spectrum and pathophysiology of TSS have not been fully elucidated.[4] These facts support the need for a method of characterizing TSS S. aureus strains.

II. OBJECTIVES

The main objectives of this project were:

(1) To investigate methods and technologies which may be used to detect and identify extracellular marker proteins (toxins) elaborated by strains of S. aureus associated with clinical cases of TSS.

(2) To examine for marker proteins four isolates of S. aureus cultured from different anatomic sites of a single individual with clinical TSS.

III. GROWTH MEDIUM:

Several attempts were made to produce a medium which would support the growth of S. aureus and induce the production of the protein marker most often seen in TSS strains, pyrogenic exotoxin C (PEC).[9, 10]

The medium must have several unique characteristics. First, since PEC is a protein, the medium should be free of proteins and polypeptides larger than 22,000 molecular weight (the estimated mol. wt. of PEC). This is to prevent interference by non-exotoxin proteins in the

isoelectric focusing procedure. Secondly, the medium must support the growth of S. aureus such that the organism reaches the log phase of growth quickly and is provided with a non-protein source of nutrition for the first 16 h of growth. This is necessary to insure that the organism is not required to metabolize proteins, i.e., PEC, for energy. Thirdly, the medium and conditions of growth should encourage the organism to produce PEC.

The most satisfactory medium was a modified version of that used by the Centers for Disease Control.

(1) Medium preparation: (See Appendix A).

IV. EXTRACTION OF PEC FROM GROWTH MEDIUM:

Two different extraction procedures were investigated. Schlievert et. al. (1977) described a method for extracting group A streptococcal pyrogenic exotoxin C.[11] Barbour (1981)[12] used a slightly different method for extracting PEC from S. aureus strains. Both of these methods are based on the selective precipitation of PEC by absolute ethanol at -20°C, and its solubility in water. We found the procedure of Schlievert to be more reproducible and required less time than that of Barbour.

(1) Extraction procedure: (See Appendix B).

V. ISOELECTRIC FOCUSING ELECTROPHORESIS:

Isoelectric focusing electrophoresis (IFE) was used to identify extracts containing PEC. Since PEC is known to have an isoelectric point (pI) of 7.2 it was necessary to utilize a technique which would allow proteins to focus at their exact pIs. IFE has several advantages

over standard electrophoresis procedures. First, it is considerably more sensitive in detecting subunit proteins. Proteins with pI differences of 0.0025 pH units can be resolved by IFE. Secondly, the procedure can be accomplished in just a few hours since focusing is carried out at high voltage.

The LKB model 2103 IFE unit was used to carry out all electrophoresis experiments in this study.

(1) IFE procedure: (See Appendix C).

VI. STAINING OF IFE GELS:

After IFE, all gels were fixed and stained with Bio Rad Silver Stain.[13] We found this stain to be far superior to Coomassie Brilliant Blue, used by some investigators in that it detected protein bands which were not detected by Coomassie Brilliant Blue. This was especially true of those weak PEC producing strains of S. aureus. Most of the proteins in the extracts migrated toward the anode and were in relatively high concentrations. These bands were readily stained by Coomassie. However, those proteins, including PEC, which focused near the cathode were few and often in low concentrations. It is our opinion that the use of Coomassie Brilliant Blue can result in an error in low PEC producing S. aureus strains.

(1) Staining procedure: (See Appendix D).

VII. RESULTS

(A) Four isolates of S. aureus from different anatomic sites of a single individual with clinical TSS were examined for PEC. It is not known if individuals with TSS must have multiple foci of lesions infected

with toxin producing S. aureus, or if multiple isolates of S. aureus from a single individual with TSS can be quite different genetically and only one be a toxin producer. This study was an effort to answer that question in one case.

Of the four isolates, two were from urine specimens collected at different times, one was from the cervix and a fourth was cultured from the throat. These isolates were compared previously using several parameters other than PEC production.

(1) Phage Typing: At RTD, the four isolates were non-typable; however, the two urine isolates showed weak (±) lytic reactions with group II typing phages 3A and 3A/71, respectively. When typed at 100 X RTD, these two strains gave strong (50 plaques to confluent lysis) reactions with typing phage 71 and weaker (+ to ±) reactions with typing phages 3A/55 and 3A/3C/55, respectively. These typing patterns may be considered similar since weak reactions at 100 X RTD are not generally considered in determining relatedness. The Staphylococcus isolate from cervical exudate remained non-typable at 100 X RTD, while the isolate from the throat gave a strong reaction with typing phage 71 and a weaker reaction with typing phage 3A. On the basis of these findings, at least three of the four strains (urine X2 and throat) appeared related due to their strong typing reactions with the Group II phages.

Further comparison of these four strains using plasmid analysis indicated that the two strains from urine had the same profile. A single plasmid DNA band was noted for each strain, and the migration distances are equivalent suggesting they are identical plasmids.

Antibiotic susceptibility patterns of each of the S. aureus isolates were similar and showed resistance to penicillin and ampicillin. None of the strains produced beta lactamase. The minimum inhibitory concentration for penicillin was 8 and 64 ug/ml for two S. aureus isolates from urine, 16 ug/ml for the throat culture isolate, and 128 ug/ml for the cervical isolate. All four strains were examined for PEC as previously described. Two positive controls and a hemoglobin standard (containing A, F, S and C hemoglobins) were also included on the gel. Silver staining revealed a protein band at pH 7.2 characteristic of pyrogenic exotoxin C only with the cervical isolate and the postive controls (Fig. 1).

This suggests that the three non-cervical isolates are different from the cervical isolate. It further shows that multifaci of toxin production are not necessary to induce TSS and, indeed, may rarely occur.

(B) Four isolates of S. aureus randomly selected from burn patients at Brook Army Medical Center were examined for PEC. All of the strains were of a common phage type (47/54/75). None of the patients exhibited any signs of TSS. The purpose of this investigation was to determine if random isolates involved in non-cervical infections produced PEC. None of these isolates were found to produce the characteristic protein band at pH 7.2.

Hemalysis patterns described by Barbour[13] were performed on all of the PEC-positive and PEC-negative isolates, but were found to be inconclusive as an indication of TSS strains.



Figure 1. IFE gel stained with Bio-Rad Silver showing PEC Bands at pH 7.2 on positive controls (189 & 033) and cervical isolate (130). HgB = hemoglobin control; 189 and 033 = positive controls; 130 = cervical isolate; 129 and 130 = urine isolate; 132 = throat isolate; pH = pH gradient.

VIII. RECOMMENDATIONS:

Future studies should be carried out to determine the immune status of individuals recovered from TSS and those of the general population. This will require the isolation and purification of PEC to use as antigen in *in vitro* testing.

Other areas of future investigation should be directed toward producing and purifying antibodies to PEC. These specific antibodies could then be used to develop a rapid screening test for PEC.

A more practical and reproducible test must be developed before TSS strains can be identified routinely in the clinical laboratory.

REFERENCES

1. Centers for Disease Control "Toxic-shock syndrome-United States, 1970-1980". Morbid. Mortal. Weekly Rep. 1981. 30: 25-28, 33.
2. Davis, J.P., P.J. Chesney, P.J. Wand, M. La Venture and the Investigation and Laboratory Team. "Toxic-shock syndrome: epidemiologic features, recurrence, risk factors, and prevention" N. Engl. J. Med. 1980. 303: 1429-1435.
3. Shands, K.N., G.P. Schmid, B.B. Dan, D. Blum, R.J. Guidotti, N.T. Hargrett, R.L. Anderson, D.L. Hill, C.V. Broome, J.D. Band, and D.W. Fraser. "Toxic-shock syndrome in menstruating women. Associated with tampon use and Staphylococcus aureus and clinical features in 52 cases" N. Engl. J. of Med. 1980. 303: 1436-1442.
4. Davis, J.P., M.T. Osterholm, C.M. Helms, J.M. Vergeront, L.A. Wintermeyer, J.C. Forfang, L.A. Judy, J. Rondeau, W.L. Schell, and the investigating team. "Tri-state toxic-shock syndrome study. II Clinical laboratory findings". J. of Inf. Dis. 1982. 145: 441-448.
5. Todd, J., M. Fishaut, F. Kapral, and T. Welch. "Toxic-shock syndrome associated with phage-group-I staphylococci". Lancet. i i: 1116-1118.
6. Aranow, H., Jr., and W.B. Wood, Jr. "Staphylococci infection simulating scarlet fever." J. Am. Med. Assoc. 1942. 119: 1491-1495.
7. Dunnet, W.N., and E.M. Schallibaum. "Scarlet fever-like illness due to staphylococcal infection." Lancet. 1960. ii: 1227-1229.

8. Stevens, F.A. "The occurrence of Staphylococcus aureus infection with scarlatiniform rash." J. Am. Med. Assoc. 1927. 88: 1957-1958.

9. Schlievert, P.M., K.N. Shands, B.B. Dan, G.P. Schmid, and R.D. Nishimura. "Identification and characterization of an exotoxin from Staphylococcus aureus associated with toxic-shock syndrome". J. of Inf. Dis. 1981. 143: 509-516.

10. Freely, J., Personal communication.

11. Schlievert, P.M., K.M. Beltin, D.W. Watson. "Purification and characterization of group A streptococcal pyrogenic exotoxin C." Inf. and Immun. 1977. 16: 673-679.

12. Barbour, A.G. "Vaginal isolates of Staphylococcus aureus associated with toxic-shock syndrome." 1981. Inf. and Immun. 1981. 33: 442-449.

13. Bio-Rad Bulletin 1089. 1981.

Appendices

A. Preparation of Fresh Beef Heart Medium for the Production of PEC.

1. Obtain several pounds of fresh beef heart. Defat and cube.

2. Add 400 ml of pyrogen-free water for each 1 lb. of beef heart and grind in a Waring Blender until a near-liquid solution results.

3. Add 1.6 g Difco trypsin (1:250) for each 1 lb. ground beef heart.

4. Adjust pH to 8.0 by adding 2.5 N NaOH slowly.

5. Place flask with beef heart in a 40°C water bath. Stir mixture often.

6. Maintain pH of the mixture at 8.0 by frequently adding 2.5 N NaOH.

7. When the pH stabilizes at 8.0, the digestion is completed-should require about 1.5-3.0 hr.

8. Dialyze mixture by placing digested heart into pre-rinsed dialysis tubing (12,000-14,000 MWCO), and dialyze against pyrogen-free water (2 liters water to 1 lb. of heart).

9. Dialyze for 72 hrs. at 4°C. Do not stir.

10. Remove dialysis tubing containing beef heart and discard. Save dialyzate and autoclave at 121°C for 15 min. Cool and refrigerate until use.

11. Add 5 ml buffer* per 100 ml medium immediately prior to use.

*Buffer:

Glucose	6 g
$Na_2HPO_4 \cdot 7H_2O$	3 g
NaCl	4 g
$NaHCO_3$	4 g
L-gluramine	0.4 g
pyrogen free H_2O	100 ml

Filter sterilize with 0.45 micron filter and store at 4°C.

B. Extraction of PEC from Growth Medium.

PROCEDURE:

1. Heavily inoculate the suspect S. aureus strain being tested for PEC into 100 ml of fresh beef heart medium containing 5 ml buffer. (Add sterile IN HCL to the broth after the buffer has been added to bring the pH down to approximately 7.6. This usually requires about 1.0 ml, but it must be determined for each batch of medium.) The S. aureus isolate must be a pure culture and actively growing on blood agar.

2. Incubate the inoculated broth at 37°C for 12-15 hours in the Psychrotherm Incubator at 60 rpm. Cultures inoculated at 1530 hr and harvested at 0800 the following day have been found to be satisfactory.

3. After removing culture from the incubator, check pH of broth with pH meter or pH paper. Adjust to 7.6-8.0 with IN NaOH. Do not assume that every culture can be adjusted by adding the same amount of NaOH to each. Each isolate will vary depending on the original inoculum size and its capacity to utilize glucose. Failure to adjust the pH at this point will result in loss of PEC. Toxin will centrifuge-out with the cells.

4. Remove cells by centrifuging the broth for 20-30 mins at 16,000 X g and 10°C (9000 rpm using 5.75 inch radius head on Sorvall RC-5B Centrifuge).

5. Supernate should be poured into 1000 ml beaker which has been placed in a wet ice bath and cooled at -20°C for 30 mins. Discard the pellitized cells.

6. Precipitate the PEC by slowly adding 400 ml of cole (-20°C) absolute ethanol to the supernate while constantly stirring with a sterile

1.0 ml pipette or glass rod. Replace the beaker in the -20°C freezer for 1 hour. (Do not remove from the wet ice bath.)

7. Separate the PEC by centrifuging at 16,000 X g for 20 mins in refrigerated centrifuge. (11,500 rpm using 4.25 in radius head on the Sorvall RC-5B Centrifuge) At this point, it is important to make sure the centrifuge tubes are free of water since the toxin is water soluble. This can be accomplished by rinsing with a small quantity of absolute alcohol. If you are certain that your final ethanol concentration is above 70% (v/v) the alcohol rinse step probably isn't necessary.

8. Pour off alcohol/broth supernate and discard. Add a small quantity of previously cooled (-20°C) ethanol to each centrifuge tube and dislodge the precipitate from the centrifuge tube with a 10 ml pipette. (It is best to score a disposable 10 ml pipette about 1/2 inch from the tip with a knife or scapel blade and break-off the tip so that the hole in the tip of the pipette is made larger. This helps in getting the precipitate out of the tubes.) Pool the precipitate of each isolate into one tube and recentrifuge for 10 mins at 16,000 xg. Again, be certain that the centrifuge tubes are free of moisture.

9. Pour off supernate and discard. Add 1.0 ml of pyrogen-free water to the precipitate. Using a sterile 1.0 ml glass pipette, mix the water and precipitate well, making an effort to break the precipitate as much as possible.

10. Remove any undissolved precipitate by centrifuging again at 16,000 xg for 15 minutes.

11. Carefully aspirate the supernate with a pasteur pipette and place into a dialysis tube (12,000-14,000 MWCO, 6mm dia.) and dialyze overnight against pyrogen-free water. This will remove any excess salt that will produce "fuzzy" bands in the isoelectric focusing step.

12. Store extract in refrigerator at $4^{o}C$ if isoelectric focusing is to be done within a few days: otherwise store at $-20^{o}C$ until ready to use.

C. Isoelectric Focusing Electrophoresis of PEC.

PROCEDURE:

1. Turn on refrigerated, circulating water bath; make sure the temperature is stable ($\pm$ 1°C) at 10°C. This usually results in a surface temperature on the gel plate of about 15°C.

2. Cover plate with a moderate layer of light mineral oil. Place the template on the plate being careful not to trap air bubbles between the template and the plate.

3. Cover the upper surface of the template with a moderate layer of light mineral oil.

4. Remove the PAG plate (pH gradient 3.5-9.5) from the sealed plastic container by cutting around the edge of the gel with sharp scissors, being sure not to damage the gel. Leave the polyester gel support and the polyester covering on the gel. Apply the gel to the upper surface of the plate making sure not to trap air bubbles between the template and the polyester gel support. At this point you should have a layer of mineral oil, a template, another layer of mineral oil, and then the PAG plate. Once the air bubbles are teased out, the upper polyester covering can be carefully removed. Leave the surface of the gel exposed for about 10 minutes while you are applying the electrode icks.

5. Remove two of the wicks from the plastic bag. Saturate the strip to be placed on the cathode side (-) with 1M NaOH. Saturate the anode strip with 1M H_3PO4. Lay the strips in the correct positions along the sides of the gel.

6. Place two small drops of sample from a vertically held pasteur pipette on the number 1 spot of the sample application area of the gel. If the samples are applied below their pI (isoelectric point) the protein will often precipitate on the gel resulting in a negative test.

If you are using a hemoglobin control (hgbs A, S, C, F), apply it at the number 7 sample application position. It is advisable not to apply a sample within the first and last inch of the PAG plate. Complete the sample and control application step as quickly as possible making sure that the samples do not run together, and that their paths do not overlap.

7. Carefully place the isoelectric focusing lid containing the electrodes on the gel. Attach the connecting wires for each electrode. See that the electrode surfaces make contact with the wicks all along the wick surface.

8. Place the cover on the electrophoresis cell securely.

9. Set the LKB 2103 power supply as follows: Set power to 30, current MA to 5, ampmeter to A 10X, and current toggle switch to 10, 10X.

10. Turn on main power switch. Slowly turn voltage dial clockwise until 79 is reached. At this point, you will notice that the current mA indicator light is on. You will also notice that KV and mA gauges are indicating low voltage and relatively high amperes. As the pH gradient is established by the ampholytes, the amperes will decrease and the voltage light will go on. (The high voltage red indicator light

will probably remain on during this procedure). The change in voltage and amperes is necessary in order to maintain constant power at 30 watts ($P = VxC$).

NOTE: The major advantages of isoelectric focusing are that a pH gradient is established in the electrophoresis medium (gel), and that high voltage can be applied across the gel since the gel is cooled by the circulating water. Also, the protein bands are not as likely to move from their pI as with ordinary electrophoresis. However, samples should not be electrophoresed too long because some drift will occur.

11. Allow electrophoresis to occur for 1.5 hr. Turn the main power switch to off before removing the cover. Set up and standardize a pH meter next to the electrophoresis cell. The meter must be equipped with a surface electrode preferably one with a small contact surface, i.e., 5 mm. Check the pH of the gel surface at 1 cm intervals along the width of the gel. It is best to do this at or near the end where no sample has been applied.

Plot a gradient curve with the cm along the X-axis and the pH along the y-axis. Graft sheets are furnished with PAG plates. Draw graft from the cathode (-) side to the anode (+) side. Usually, once a curve has been drawn for a given set of conditions, it is not necessary to do this with each run. However, should you change the procedure in any way, a new pH gradient curve must be established.

12. Gently remove the gel from the plate being careful not to touch the gel surface. Place gel in stain fixative.

D. Staining of IFE Gels:

1. Fix in 30% methanol/10% TCA/3.5% SSA for 1 hour. (75 mls methanol/ 25g TCA/8.75g SSA; q.s. to 250 mls with deionized water)

2. Transfer thru several changes of 30% methanol/12% TCA to remove the ampholytes: this should take at least 2 hours. (225 mls methanol/ 90g TCA; q.s. to 750 mls with deionized water.)

3. Continue fixing in 10% ETOH/5% HAC thru two changes; 30 mins each. (50 mls ETOH, 25 mls HAC; q.s. to 500 mls with deionized water.)

4. Place gel in oxidized solution for 10 mins. (40 mls oxidizer solution; q.s. to 400 mls with deionized water.)

5. Wash three times in deionized water; 10 mins each time. (Two times for 5 mins each if gel has a polyester backing, i.e., PAG Plates. Gel should be pale yellow and clear.)

6. Transfer gel to silver reagent for 30 mins. (40 mls silver reagent; q.s. to 400 mls with deionized water.)

7. Wash gel in deionized water for 2 mins. (1 min for polyester-backed gels.)

8. Place gel in developer solution for 1 min. This step is to remove excess silver deposits from the gel surface. (Dissolve 32g of developer crystals in 1 liter deionized water.)

9. Transfer gel to fresh developer solution for approximately 5 mins. Watch bands closely during the developer steps to insure that the staining intensity is optimal for your needs.

10. Transfer gel to Stop Solution for 5 mins. (25 mls HAC; q.s. to 500 mls with deionized water.)

NOTE: All stock reagents should be kept at 4 C until ready for use. Oxidizer and silver working reagents should be used only once: developer may be left at room temperature for 1 month before use.)

TCA = Trichloroacetic acid

SSA = Sulfosalicylic acid

HAC = Acetic acid

ETOH = Ethanol

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

EFFECT OF NDFDA ON TISSUE FATTY ACIDS OF THE RAT

Prepared by:	Dr. Thomas E. Webb
Academic Rank:	Professor
Department and University:	Department of Physiological Chemistry, College of Medicine; The Ohio State University
Research Location:	Air Force Aerospace Medical Research Laboratory, Toxic Hazards Division, Toxicology Branch
USAF Research Colleague:	James P. Murphy/Dr. Melvin E. Andersen
Date:	August 2, 1982
Contract No:	F49620-82-C-0035

EFFECT OF NDFDA ON FATTY ACID SYNTHESIS IN RAT LIVER

by

Thomas E. Webb

ABSTRACT

The question of the biochemical basis for the extreme hepatotoxicity of NDFDA (nondecafluorodecanoic acid) has been investigated from the point of view of its effects on stearoyl CoA desaturase and associated electron transport functions in the microsome fraction. Changes in this parameter would be consistent with perturbations observed earlier in the fatty acid content of liver and of membranes of cells exposed to this compound. These preliminary studies indicate that hepatic stearoyl-CoA desaturase activity drops in both NDFDA-treated animals and in the pair-fed controls, approaching zero within 6-8 days. This drop is attributed mainly to decreased food intake and since it occurred also in the controls, does not appear to be related to NDFDA hepatotoxicity. However, NDFDA induced several biochemical changes not present in the pair-fed controls. These include (1) decreased inducibility of stearoyl-CoA desaturase by force feeding an amino acid/sucrose mixture, (2) a marked decrease in the rate of microsomal electron transport from NADH through cytochrome b_5 to the terminal oxidases (including the desaturase) and to molecular oxygen and (3) an increase in the concentration of cytochrome P-450, an important component of the microsomal drug metabolizing system. NDFDA caused only a small decrease in the content of microsomal cytochrome b_5. Insulin levels appeared normal in the treated animals. The unique effects of NDFDA on one of the main target organs appears to reflect direct or indirect modulation of genetic expression and/or specific changes in membrane systems rather than non-specific cellular damage.

ACKNOWLEDGEMENT

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to spend a worthwhile and interesting summer at the Air Force Aerospace Medical Research Laboratory, Wright Patterson Air Force Base, Ohio. He would like to acknowledge the laboratory, in particular the Toxicology Branch for its hospitality and excellent working conditions.

Finally he would like to thank Dr. Melvin E. Andersen for suggesting this area of research and would like to thank James P. Murphy for his collaboration, assistance and guidance. He would also like to acknowledge many helpful discussions of Ms. M. E. George, Dr. C. T. Olson, Dr. W. C. Keller and Mr. M. L. Gargas as well as assistance of other members of the Toxic Hazards Section.

I. INTRODUCTION

NDFDA (nondecafluorodecanoic acid), a straight chain perfluorinated 10 carbon acid ($CF_3(CF_2)_8\ CO_2H$) is similar in structure to compounds used as aqueous film forming foam fire extinguishants. These polyfluorinated chemicals are also used to impart oil and water resistance to porous materials. NDFDA in particular is extremely toxic and may serve as a model compound for evaluating the health hazards associated with the use of these compounds in the Air Force and other Armed Forces and by the general public at large. In order to evaluate the health hazards of NDFDA it is necessary to elucidate the mechanism of its toxicity. Particularly appropriate is an evaluation which will permit the development of appropriate diagnostic and therapeutic methodology for short duration-high level infrequent exposure[1].

Toxicity studies to date on NDFDA have been carried out almost exclusively at the AFAMRL. These studies [2-5] indicate that dosages in the region of the LD_{50} ie. 50 mg/kg body weight, causes rats to become anorectic and to undergo a rapid weight loss. The main pathological changes occur in the liver, both the nucleus and cytoplasm undergoing change. (The bone marrow, thymus, stomach and testes are also affected.) Changes have been shown to occur in the properties of the plasma membranes of cells exposed to NDFDA[6] and significantly changes also occur in the concentration of several fatty acids in the liver[5] which serve as precursors for membrane lipids and phospholipids. More specifically there is a marked increase in the oleic acid content of the liver with a decrease in the hepatic concentration of stearic acid, its immediate precursor. The oleate levels remain elevated for as long as 30 days; this change was not observed in pair-fed controls.

The identification of NDFDA-induced alterations in hepatic levels of stearic and oleic acid, together with evidence that this is one of the main target organs for NDFDA and indications from other studies that membranes from NDFDA exposed cells are altered, focused attention on reactions catallyzing the conversion of stearic acid to oleic acid. Of particular interest was the effect of NDFDA treatment on stearoyl CoA

desaturase, the enzyme which carries out the desaturation of stearic acid to oleic acid. This reaction, which occurs in the microsome fraction of the liver requires NADH and molecular oxygen as cofactors and the CoA thioester of the fatty acid as substrate. As indicated in Fig. 1, this reaction requires a multicomponent system[7,8] which includes 3 microsomal electron transport proteins. A number of studies to date indicate that stearoyl-CoA desaturase activity is regulated by the amount of terminal oxidase, which has a short half-life (4 hours) as compared to cytochrome b_5 (120 hours) and cytochrome b_5 reductase (144 hours). Its activity drops drastically on fasting[7,8] and is very low in diabetic rats[9]. It is induced to levels several-fold higher than present in the livers of fed rats when fasted rats are refed on fat-free high carbohydrate diet, or diabetic rats are treated with insulin.

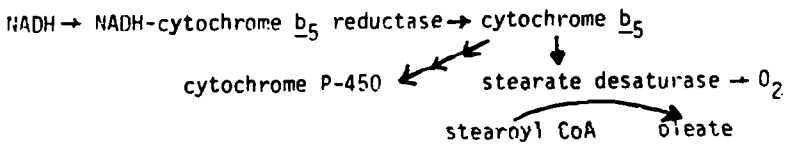


Fig. 1. Pathway for microsomal electron transport from NADH to the terminal oxidase "desaturase" and in the absence of stearate (as the CoA thioester) or other fatty acid substrates, to cyt. P-450 or other terminal oxidases.

II. OBJECTIVES

The overall objective of this study is to investigate the biochemical mechanisms underlying the toxicity of NDFDA on one of the main target organs-the liver. In view of documented pertabations in the fatty acid content of livers of NDFDA-treated rats and related evidence for changes in membrane properties, the specific objective was to measure in the microsome fraction from the livers of NDFDA-treated rats, their pair-fed controls, and normal (ad libitum-fed) rat, the following parameters:

(1) The overall activity of stearoyl Co-A desaturase for reasons outlined above.

(2) To assess the status of individual components of the desaturase system including cytochrome b_5 and the desaturase (terminal oxidase) and to assess the status of cytochrome b_5 reductase by indirect measurements.

(3) To assess the status of cytochrome P-450 which probably serves as a terminal oxidase (electron acceptor) for reduced cytochrome b_5 in the absence of stearoyl -CoA.

(4) To assess the insulin status of the rats since the desaturase is depressed in diabetic rats and is induced by insulin[8].

(5) To compare the inducibility of the liver desaturase by dietary means.

It should be noted that this investigation assessed not only the effect of the toxic compound on fatty acid metabolism, but also on microsomal electron transport functions and a main component (cytochrome P-450) of the microsomal drug metabolizing system in the liver. It also recognizes that the level of hepatic stearoyl CoA desaturase activity (more specifically the terminal oxidase) is sensitive to this hormonal and nutritional status of the animal and includes appropriate controls. It also tests the effect of NDFDA on the inducibility of the enzyme, which serves as a measure of the intactness of the regulatory controls in the rat.

III. <u>The Level of Stearoyl CoA-Desaturase Activity and its Inducibility in the Liver Microsomes of NDFDA-treated Rats and in their Pair-fed Controls</u>.

These experiments were carried out on <u>ca</u> 200 gm male rats of the Fischer strain which were fed rat chow. Since NDFDA-treated rats drastically curtail their food intake, untreated pair-fed controls were run in parallel. In a typical experiment 8-10 rats were injected ip with a single dose of 50 mg/kg of NDFDA in ethylene glycol: water (1:1, v/v) as vehicle, then 2 animals were sacrificed on each of days 2, 4, 6 and 8. Pair-fed controls injected with the vehicle on day 1 were also sacrificed

on alternate days. Body weights and food consumption were recorded daily. The surgical removal of the liver, the drawing of blood by cardiac puncture and force feeding of the animals by stomach tube was performed, where indicated, while the animals were under halothane anesthesia. Although halothane is known to enhance electron transfer from cytochrome b_5 *in vitro*[11], mM amounts are required, i.e. much more than that accumulating during anesthesia.

The microsome fraction was purified from the livers and analyzed for stearoyl-CoA desaturase activity as described by Strittmatter *et al*[7]. This method is based on isolation by differential centrifugation and includes several washings in the ultracentrifuge to free the microsomal membranes of absorbed contaminating proteins and substrates. The spectrophotometric assay was adapted to a Gilford Model 2600 spectrophotometer, which is microprocessor controlled and equipped with a plotter. The assay involves the detection of the onset of cytochrome b_5 reoxidation in a microsome suspension at 424 nm and 30°C after reduction with 1.0 n mole of NADH in the absence (B) and presence (A) of excess stearoyl CoA as shown by the tracings in Fig. 2. Cytochrome b_5 is very rapidly reduced by NADH through the mediation of cytochrome b_5 reductase and remains reduced until all of the NADH is oxidized by electron transfer through b_5 to oxygen by some autooxidation, but mainly by transfer to terminal oxidases such as cytochrome P-450[10]. Once all NADH is oxidized cytochrome b_5 becomes oxidized in turn and the A_{424} decreases. In the presence of stearoyl CoA electrons are transferred from cytochrome b_5 to the desaturase and NADH is more rapidly oxidized. Thus assuming 1.0 mole of NADH is required for each mole of oleyl-CoA formed the activity in nmol/min/mg protein=

$$\frac{B-A}{B \times \text{mg protein}}$$

where B and A = time for reoxidation of NADA in absence and presence of stearoyl CoA respectively.

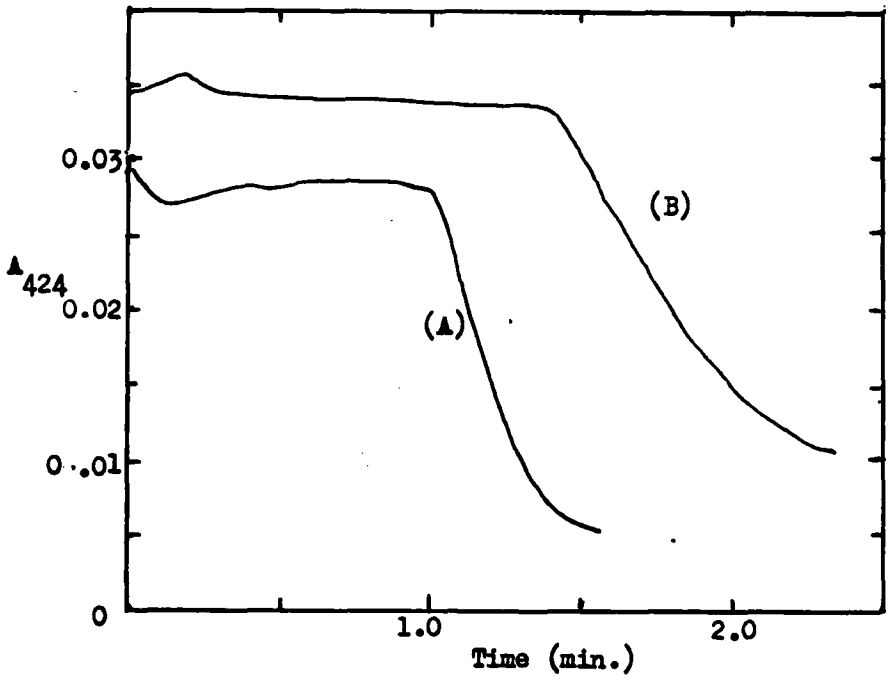


Fig. 2. Change in A_{424} as a function of time during the reoxidation of 1nmole of NADH by a microsome suspension from a normal red rat in the absence (B) and presence of excess stearoyl CoA. In this example A and B were estimated to be 1.425 and 1.00 min, respectively.

A summary of the changes in stearoyl CoA desaturase activity in NDFDA-treated animals and their pair fed controls as a function of time post-treatment and the inducibility of the enzyme by dietary means, is shown in Fig. 3a and b. The data indicate that the enzyme decreases dramatically in both groups of animals, the activity approaching zero within 4-6 days. It is concluded that most of the decrease can be attributed to decreased food intake; the latter decreased roughly linearly in the NDFDA-treated rats and was essentially zero by 6 days post-treatment.

In contrast a real difference was observed between the NDFDA-treated rats, and their pair-fed controls when enzyme inducibility was assessed. In these experiments the treated or pair-fed rats were force-fed (intubated by stomach tube) with 5.0 ml of a aqueous solution (pH 7.0) of 20% sucrose and 20 gm % Bactotryptone (a pancreatic casein digest) 15 hours before sacrifice. In the pair fed controls the fasting imposed by the pair feeding regimen acted in concert with the 15 hour forced feeding to induce the desaturase activity to levels twice those found in the liver of a normal fed rat. In contrast only slight induction was observed during the earlier phase of the experiment in the livers of NDFDA-treated rats. Induction appears to be absent 8 days after treatment.

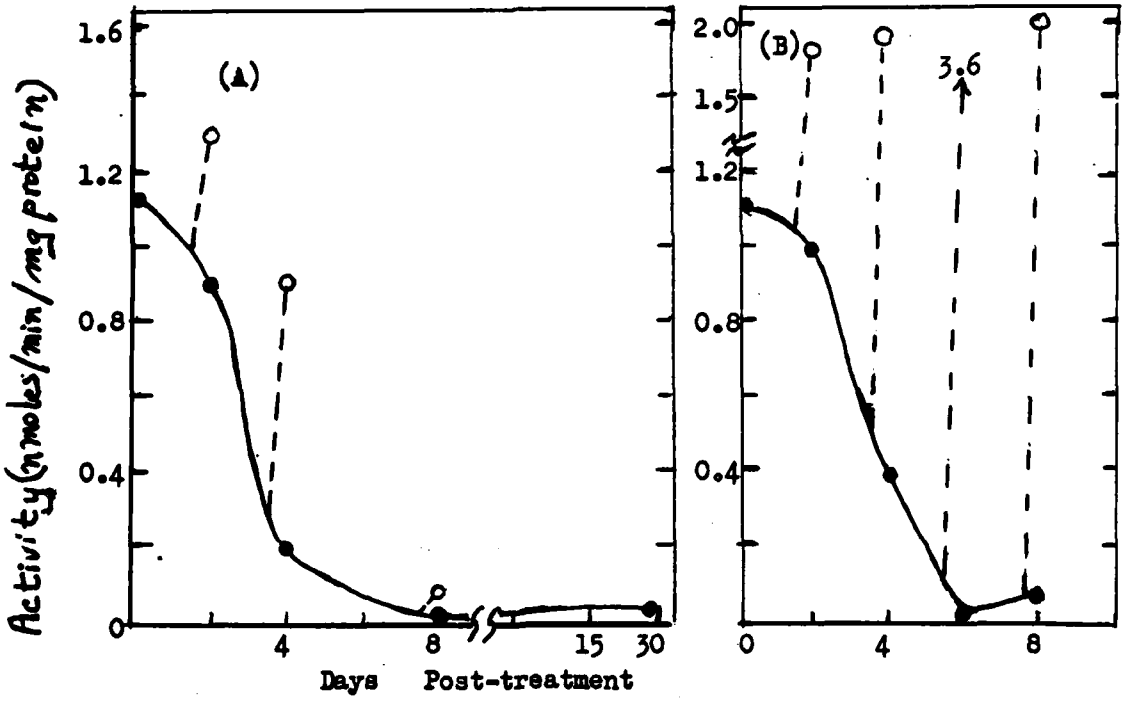


Fig. 3. Time-course changes in the activity of stearoyl CoA desaturase activity in the livers of (a)rats treated with NDFDA or (b) pair-fed controls as a function of duration of post-treatment (●—●). The 15 hour time course of induction in response to forced feeding is also shown (o---o). In some instances where there was no stearoyl CoA desaturase activity in the microsomal preparation, the addition of stearoyl CoA actually delayed further the oxidation of NADH.

Since stearoyl CoA desaturase activity is inducible by insulin and indeed may require the presence of insulin for dietary induction, the insulin status of these animals was assessed by measuring their blood glucose levels. Blood glucose was measured in serum samples from selected rats by the Pathology section of the Toxicology Division. The blood glucose levels on individual rats are tabulated in Table 1. It is concluded that the blood glucose levels are not significantly higher in the NDFDA-treated rats indicating that insulin levels are within the normal range in these animals.

TABLE I

Serum Glucose Levels in Normal and NDFDA-Treated Rats

Serum Donor	Serum Glucose * mg/dl
1. Normal fed rat	151, 137
2. NDFDA-treated rat 8d post-treatment	118, 142, 136
3. (2) 15 hrs after force-feeding	92, 117, 125

* Individual values are shown.

IV. Effect of NDFDA on the Rate of Microsomal Electron Transport from NADH to Molecular Oxygen.

As indicated in Table II pretreatment of rats with NDFDA resulted in a significant (2-3 fold) drop in the rate of reoxidation of NADH by the microsome fraction in the standard stearoyl CoA desaturase assay; this delay was observed both in the absence and presence of stearoyl CoA. However, the rate of reoxidation of NADH by microsomes from the pair-fed controls was similar to that of microsomes from normal fed rats. Force-feeding of the animals did not alter these differences. Since the reduction of cytochrome b_5 by NADH appeared to be very rapid in all assays (ie. was not observable under our conditions of assay) the cytochrome b_5 reductase appears to be present in excess in both the NDFDA-treated and pair fed animals. The lesion appears to be distal to cytochrome b_5, the oxidation of which is also unchanged. Further, since it is observed in both the presence and absence of stearoyl CoA it must be proximal in the pathway to the terminal oxidases.

In some instances where there was no stearoyl CoA desaturase activity in the microsomal preparation, the addition of stearoyl CoA actually delayed further the oxidation of NADH.

TABLE II

Time for Oxidation of NADH by Microsomes from NDFDA-treated, Pair-fed and Normal Rats

Liver donor	Time for oxidation of 1 nmole NADH (min/0.3 mg prot.)	
	NADH only	NADH + Stearoyl CoA
(1) Normal fed rats	1.97 + 0.20	1.30 + 0.15
(2) 6-8 day NDFDA treated rats	4.28 + 0.35	4.09 + 0.38
(3) (2) force fed	4.59 + 0.33	4.50 + 0.64
(4) 6-8 day pair fed rats	1.31 + 0.30	1.24 + 0.11
(5) (4) force fed	1.28 + 0.15	1.03 + 0.11

V. Effect of NDFDA Treatment on the Microsomal Cytochrome b_5 Content of Rat Liver.

The liver microsomal cytochrome b_5 content is summarized in Table III as % of cytochrome b_5 in liver microsomes from a normal fed rat. The concentration was estimated from the decrease in A_{424} incident to the reoxidation of NADH-reduced cytochrome b_5 in the standard desaturase assay (of Fig 1). An extinction coefficient (E_{mM}) of 100 was used in these calculations[12] and the values were corrected for variation of the protein content of the samples.

In general there was no marked consistent change in the microsomal cytochrome b_5 content of liver from NDFDA-treated rats. Thus, even the lowest value observed 6 days post-treatment was still 75% of the normal control value. Further, there were no striking differences between the treated and treated/force fed groups. The decrease in cytochrome b_5 content was somewhat less in the pair fed controls and was elevated above the normal control values when these animals were force fed the sucrose:casein hydrolysate mixture. In general it does not appear that changes in this parameter can account for changes in the rate of either

oleic acid production or microsomal electron transport observed in the microsomal preparations from treated rats.

TABLE III

Effect of NDFDA Treatment of Rats on Liver Microsomal Cytochrome b_5 Content

Liver donors	Cytochrome b_5 Content (% Normal Fed Control)*
(1) NDFDA treated	
2 days	83.2
4 days	75.0
8 days	72.1
(2) NDFDA treated/force fed	
2 days	88.0
4 days	82.0
8 days	77.1
(3) Pair fed controls	
2 days	86.7
4 days	91.5
8 days	107.3
(4) Pair/force fed	
2 days	149.9
4 days	112.5
6 days	90.4

*The microsomes from the normal fed control rat had 0.974 nmoles of cytochrome b_5 per mg protein.

VI. Effect of NDFDA Treatment of Rats on the Concentration of Microsomal Cytochrome P-450 in the Liver.

Cytochrome P-450 may act as terminal oxidase for the reoxidation of NADH reduced cytochrome b_5 in the absence of stearoyl-CoA, or of stearoyl CoA desaturase[10]. The concentration of cytochrome P-450 in liver microsomes was estimated by the difference spectrum of its carbon

monoxide derivative form. In this procedure the difference spectra over the range 400 and 500 nm is obtained from two microcome suspensions treated with sodium dithionite, one of which is saturated with carbon monoxide. The nmoles of cytochrome P-450 was calculated from the relation:

$$\frac{A_{450-480} \times 1000}{91 \times \text{mg protein}}$$

The limited data in Table IV suggests that cytochrome P-450 levels in the liver are raised in response to the administration of the toxic compound. This may indicate that NDFDA induces and/or is metabolized by the microsomal drug metabolizing system. This increase is even more significant in view of the decrease in cytochrome P-450 in the microsomes of the pair-fed controls. In any case, the data appear to rule out the possibility that decreased microsomal transport in NDFDA-treated rats is due to a decrease in this terminal oxidase.

TABLE IV

Effect of NDFDA Pretreatment on Liver Microsomal Cytochrome P-450 Content

Liver donors	Cytochrome P-450 (% Normal Control)
(1) NDFDA-treated rats	
4d	178.9
6d	100.0
(2) Force fed NDFDA-Treated	
4d	212.3
6d	175.4
(3) Pair-fed controls	
4d	31.75
6d	72.98
(4) Pair-fed, force fed controls	
4d	49.1
6d	81.7

VII. RECOMMENDATIONS

The spectrophotometric method employed in this study to assay microsomal stearoyl-CoA desaturase depends upon the integrated sequential functioning of several enzymes and electron acceptor proteins which are membrane components. The results of the present study indicate that both the overall desaturase activity and the rate of microsomal electron transport are reduced in the NDFDA-treated rats. Although a reduction in the terminal oxidase of the stearoyl CoA desaturase system would be consistent with the known reduction of this enzyme on starvation in these anorexic rats, due to the complexity of the system it might still be desirable to obtain a more direct estimate of the rate of conversion of stearic acid to oleic acid *in vivo*. This should be possible by injecting labeled stearic acid via the portal vein, then after several short time intervals, isolating the liver fatty acids, and separating them prior to determining the concentrations of labeled stearic and oleic acids. Comparisons could be directly made between the NDFDA-treated rats, their paired controls and normal fed rats after correction for differences in pool size; a considerable amount of information is already available concerning the pool sizes in these 3 classes of rats.

The decreased microsomal electron transport from NADH to the desaturase and other terminal oxidases in the NDFDA-treated rat liver is indeed of interest. It does not appear to be due to either a reduction in the concentration of microsomal cytochrome b_5, or its rate of reduction. One possibility is that the rate of transport between cytochrome b_5 and the terminal oxidase (which is thought to be cytochrome P450 in the absence of substrate for the desaturase) is markedly inhibited due to the decrease in a reductase, or to changes in the fluidity of the membrane. For example, the perturbed lipid metabolism in the liver may lead to changes in the fluidity of the membrane as has been reported in different nutritional states[14]. Aside from studies on the microsomal membrane, it may be possible to rectify these defects by force-feeding diets with fatty acid content designed to normalize the ratios in the NDFDA-treated animal.

The apparent increase in the cytochrome P-450 content of NDFDA-treated rats is unexpected since there was decreased inducibility of stearoyl-CoA desaturase by hyperalimentation suggesting that the regulatory or synthetic mechanisms in the liver were damaged. Since this data was based on a minimal number of animals, additional studies are warranted. However, the depressed values in the pair-fed controls suggest that these data are valid. These data could be confirmed in vivo simply by demonstrating that the phenobarbitol-induced sleeping time is decreased in these animals. The induction of cytochrome P-450 by NDFDA would mean that this hepato-toxin is a substrate, or pseudosubstrate, for the microsomal drug metabolizing system and would indicate that it may be metabolized in the liver. The answer to the latter problem is of interest since a metabolite of NDFDA could be the actual hepatotoxin.

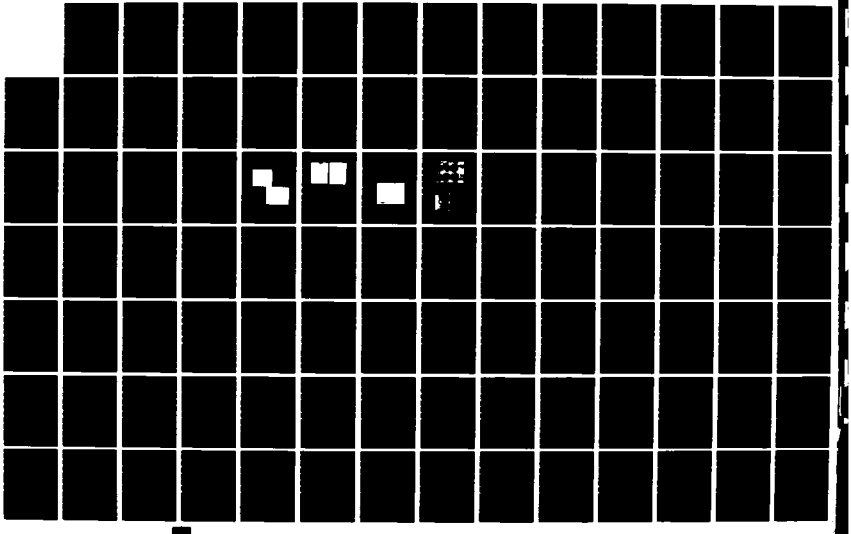
Because of the severity of the hepatotoxicity it is quite probable that liver regeneration (hypertrophy) would occur, providing the damage is not too severe. Both the surgical removal of 2/3 of the liver, or treatment with the hepatotoxin, CCl_4, are known to induce liver regeneration[12]. The relationship between liver regeneration and survivability at various dosages of NDFDA is of particular interest, since the LD_{50} may be related to the functional competence of the repairative mechanisms of the liver.

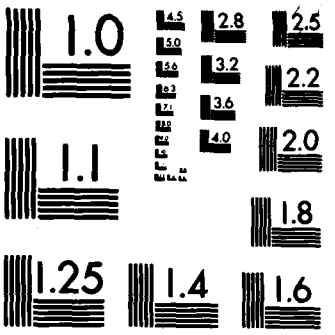
In addition to destroying liver cells, NDFDA may affect the regulatory mechanisms in the surviving cells. Although nutritional induction of stearoyl CoA desaturase is reduced, this could be the result of decreased absorption of the force-fed nutrients from the G.I. tract. The enzyme is also normally inducible by insulin and assay for induction by this hormone would be a measure of both the intactness of the genetic regulation in the cell and the status of the insulin receptors in the membrane. Another hepatic enzyme which might be tested for inducibility is tyrosine transaminanase, which is induced by insulin, glucagon and glucocorticoids[15].

The liver has been found to be protected from several hepatoxic drugs by post-treatment with chloramphenicol, an inhibitor of bacterial and mitochondrial protein synthesis[16]. Although the mechanism is not known, it would be of interest to test the effect of treating

AD-A130 770 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM (1982) RESEARCH REPORTS VOLUME... (U) SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION INC S... 10/1

UNCLASSIFIED W D PEELE ET AL. OCT 82 AFOSR-TR-83-0614 F/G 5/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

NDFDA-treated rats with chloramphenicol to determine its effect on the course of the toxic response and on survival. Finally it would be of interest to determine whether NDFDA has any carcinogenicity in addition to its hepatotoxicity, using any one or combination of presently available long-term and short-term tests.

The apparent increase in the cytochrome P-450 content of NDFDA-treated rats is unexpected since there was decreased inducibility of stearoyl-CoA desaturase by hyperalimentation suggesting that the regulatory or synthetic mechanisms in the liver were damaged. Since this data was based on a minimal number of animals, additional studies are warranted. However, the depressed values in the pair-fed controls suggest that these data are valid. These data could be confirmed in vivo simply by demonstrating that the phenobarbitol-induced sleeping time is decreased in these animals. The induction of cytochrome P-450 by NDFDA would mean that this hepato-toxin is a substrate, or pseudosubstrate, for the microsomal drug metabolizing system and would indicate that it may be metabolized in the liver. The answer to the latter problem is of interest since a metabolite of NDFDA could be the actual hepatotoxin.

Because of the severity of the hepatotoxicity it is quite probable that liver regeneration (hypertrophy) would occur, providing the damage is not too severe. Both the surgical removal of 2/3 of the liver, or treatment with the hepatotoxin, CCl_4, are known to induce liver regeneration[12]. The relationship between liver regeneration and survivability at various dosages of NDFDA is of particular interest, since the LD_{50} may be related to the functional competence of the repairative mechanisms of the liver.

In addition to destroying liver cells, NDFDA may affect the regulatory mechanisms in the surviving cells. Although nutritional induction of stearoyl CoA desaturase is reduced, this could be the result of decreased absorption of the force-fed nutrients from the G.I. tract. The enzyme is also normally inducible by insulin and assay for induction by this hormone would be a measure of both the intactness of the genetic regulation in the cell and the status of the insulin receptors in the membrane. Another hepatic enzyme which might be tested for inducibility is tyrosine transaminanase, which is induced by insulin, glucagon and glucocorticoids[15].

The liver has been found to be protected from several hepatoxic drugs by post-treatment with chloramphenicol, an inhibitor of bacterial and mitochondrial protein synthesis[16]. Although the mechanism is not known, it would be of interest to test the effect of treating

REFERENCES

1. The AFAMRL Mission
   Wright Patterson AFB, Ohio
   (U.S. Gov't Printing Office 1982
   Publication 559-007/6021)

2. Andersen, M.E., Baskin, G. and Rogers, A. "The Acute Toxicity of Perfluoro-N-Decanoxic Acid: Similarities with 2, 3, 7, 8-Tetrachlorodibenzodioxin", The Toxicologist. 1, 16, 1981.

3. Andersen, M.E., Olson, C. Rogers, A. and Van Rafelghem, M. "Toxicity of Perfluoinated Fatty Acid in Comparison to that of 2, 3, 7, 8-Tetrachlorodibenzodioxin", Presented at the Review of Air Force Sponsored Basic Research in Environmental Toxicology, The Ohio State University, Columbus, OH. 2-3 June 1981 (as cited in Ref 5, below).

4. Bacon, J., Keller, W. and Andersen, M. "Teratological Evaluation of a Model Perfluorinated Acid, NDFDA", Air Force Aerospace Medical Research Laboratory, Wright Patterson AFB, Ohio. AFAMRL-TR-81-14 (AD A095424), 1981.

5. Olson, C.T. "Effects of Nonadecafluorodecaroic Acid on Tissue Fatty Acids of the Rat." Ph.D. Disseratation, The Ohio State University, 1982.

6. Andersen, M.E. and George, M.E. Personal Communications, 1982.

7. Strittmatter, P., Spatz, L., Corcoran, D., Rogers, M.J., Setlow, B. and Redline, R. "Purification and Properties of Rat Liver Microsomal Stearyl Coenzyme A Desaturase" Proc. Natl. Acad. Sci., 71, 4565-4569, 1974.

8. Oshino, N., Imai, Y. and Sato, R. "A Function of Cytochrome b_5 in Fatty Acid Desaturation by Rat Liver Microsomes." J. Biochem. (Japan) 69, 155-167, 1971.

9. Prasad, M.R. and Joshi, V.C. "Regulation of Rat Hepatic Stearoyl Coenzyme A Desaturase." J. Biol. Chem. 254, 997-999, 1979.

10. Ivanetich, K.M., Manca, V. and Harrison, G.G. "Influence of Two Haloalkanes on the Redox Behavior of Hepatic Microsomal Cytochrome b_5 and Its Possible Relationship to Stearate Desaturase." Research Communications in Chemical Pathology and Pharmacology. 34, 473-484, 1981.

11. Ivanetich, K.M., Manca V., Harrison, G.G. and Berman, M. "Enflurane and Methoxyflurane: Their Interaction with Hepatic Microsomal Stearate Desaturase." Biochem. Pharmacol. 29, 27-34.

12. Dorman, R.V. and Webb, T.E. "Interrelationship Between Ribosome Formation and Nuclear DNA Synthesis in Regenerating Rat Liver." Experientia 30, 28-29, 1974.

13. Strittmatter, P., Rogers, M.J. and Spatz, L. "The Binding of Cytochrome b_5 to Liver Microsomes." J. Biol. Chem., 247, 7188-7194, 1972.

14. P. Mazel in Fundamentals of Drug Metabolism and Drug Disposition (B.W. LaDu, H.G. Mandel and E.L. Way eds. The Williams and Wilkins Co. 1971). pp. 573-575.

15. Kates, M. and Pugh, E.L. "Fluorescence Polarization Studies of Rat Liver Microsomes With Altered Phospholipid Desaturase Activities." Canadian J. Biochem. 58, 952-958, 1980.

16. McNamara D.J. and Webb, T.E. "A Common Component Involved in the Induction of Hepatic Tyrosine Transaminase by Hydrocostisone and Glucagon." Biochim. Biophys. Acta 313, 356-362, 1973.

17. Lubac, G.E. and Blunck, J.M. "Inhibition of Chemical Carcinogenesis." Biochem. Pharmacol. 29, 2252-2255, 1980. (See also Raisfeld, I.H. "Inhibition of β-Amanitine-induced Liver Degeneration by Chloramphenicol." in the Biochemistry of Disease, Part A, 1974, p 203.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

HOOK INTERFEROMETRY USING A PULSED DYE LASER

Prepared by:	David Weimer
Academic Rank:	Associate Professor
Department and University	Department of Physics Ohio Northern University
Research Location:	Aero Propulsion Laboratory WPAFB Aero Space Power Division Plasma Physics/POOC-3
USAF Research Colleague:	Dr. Peter Bletzinger
Date:	September 2, 1982
Contract No.	F49620-82-C-0035

HOOK INTERFEROMETRY USING A PULSED DYE LASER

by

David Weimer

ABSTRACT

The question of whether a Michelson interferometer can be used instead of a Mach-Zehnder interferometer for spectral interferogram hook measurements with a one meter grating spectrograph has been investigated. Using a pulsed He discharge and a nitrogen pumped dye laser, hooks have been measured and the population density determined. The results appear to be consistent with published data. Suggestions for further research in this area are offered.

ACKNOWLEDGMENTS

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to spend a worthwhile and interesting summer at Aero Propulsion Laboratory, Wright Patterson Air Force Base, Dayton, Ohio. In particular, I would like to acknowledge the Plasma Physics Section members for their hospitality and enjoyable working conditions.

I would like especially to thank Dr. Peter Bletzinger for his guidance and suggestions and for securing equipment for pursuing this experiment.

I. INTRODUCTION:

The hook method for the measurement of the population density of excited states has received considerable attention in recent years with the interest in laser cavity physics.

Hook interferometry is basically a tool to examine the refractive index of a gas--in particular the spectral behavior of the index. In the radiation spectrum of a gas the index changes very rapidly in the near vicinity of a line, leading to the phenomenon of anomalous dispersion. The wave length interval over which this occurs can be related to the number density in the excited state.[1,2] The method combines an interferometer with a spectrograph to spectrally analyze the fringe pattern produced by the interferometer. A broad spectrum light source is used to illuminate the interferometer, one leg of which contains the cell of test gas and the interference fringes so produced are focused on the entrance slit of the spectrograph. The result is a fringe pattern in the focal plane of the spectrograph which might be termed a spectral interferogram. At the particular wave length of a spectral line, hooks occur in the fringe pattern, and the position of these become the basic measurement. See figure 1.

The relationship between Δ and the population density is given by Huber.[2]

$$N f \ell = \pi K \Delta^2 / r_0 \lambda_0^3 \qquad (1)$$

r_0 = classical electron radius

λ_0 = wave length of the line

ℓ = optical path length

$$K = \frac{P\lambda'}{\Delta\lambda}$$

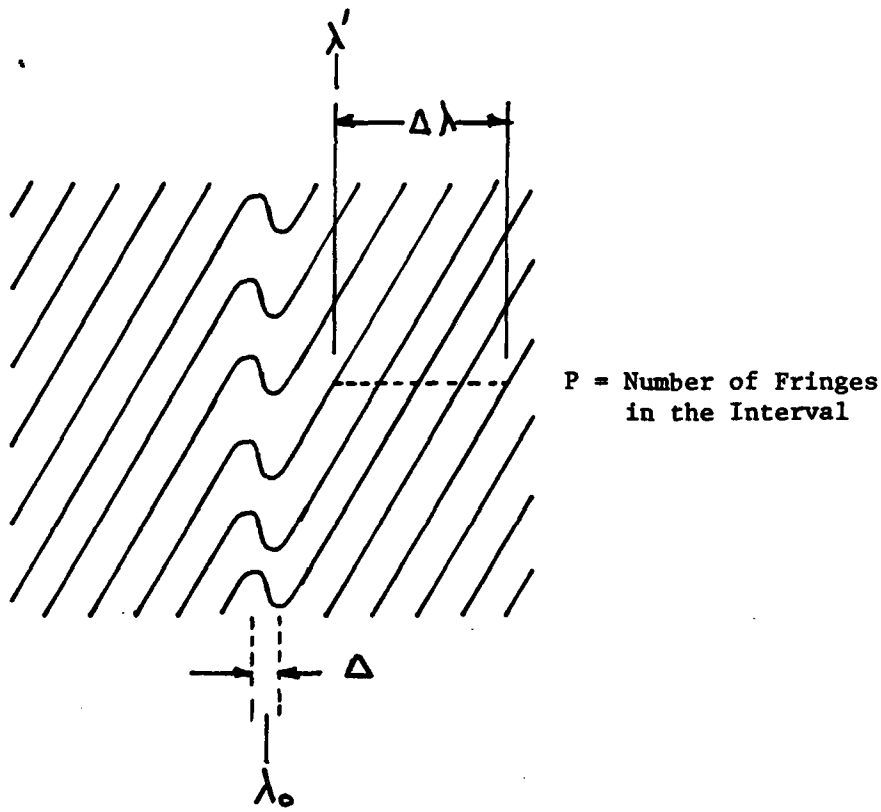


Fig. 1 Sketch of Spectral Interferogram

P = number of fringes along a horizontal line in a certain wave length interval, slightly removed from the line

$\Delta\lambda$ = length of the wave length interval

f = oscillator strength, tabulated (Wiese)[7]

The method dates back to the paper of D. Roschdestvensky in 1912 on the "Anomalous Dispersion in Sodium Vapor." [1] The technique has been applied not only to discharge tubes but to the shock heating of metals suspended in the reflected shock region of a shock tube.[2] A review paper by W. C. Marlow[3] and articles in the Darshan S. Dosanjh (Ed.)[4] book cover the background and many applications of the method. More recently nitrogen pumped dye lasers have been used with the hook method to get the time history of excited states.[5, 6]

The Plasma Physics Section of AFWAL has a particular concern in the physics of high energy discharges and an interest in developing the hook technique for diagnostic purposes.

II. OBJECTIVE OF THE RESEARCH EFFORT

The main objective of this effort was to demonstrate the feasibility of the hook technique with a helium discharge tube illuminated with a nitrogen pumped dye laser. It was proposed to use an available Michelson interferometer and a 1.0 meter Czerny-Turner Spectrograph. In the past most hook interferometers have been of the Mach-Zehnder type. It was also proposed to evaluate this latter type as well.

III. THE APPROACH

A model NRG-DL-0.03 tunable dye laser was modified by removing the echelle grating in the rear of the cavity and replacing it with a plane mirror. This provides the dye cell laser with a broad band of radiated

light to illuminate the discharge tube. The band width for the dye Rhodamine 6G ($5x10^{-3}$ moles/liter in ethonal) was from approximately 5750 A^{o} to 6050 A^{o}. The He line of interest was 5876 A^{o} which is a transition between $2^3P - 3^3D$ states. The illumination time is about 5 nanoseconds. A Cenco model M-4 Michelson interometer was used. This has an aperture of 1¼" x 1¼" with a clear space of 3½" to insert a test cell. This was then the limiting dimension for the discharge tube. The one meter spectrograph was limited to first order in this part of the spectrum and the 1180 groove/mm grating had a dispersion of 8.2 A^{o} /mm. Since it is necessary to have a resolution of at least 2 A^{o} /mm, a secondary camera to image the spectral focal plane was attached to the spectrograph. This gave a magnification of about 3.8 and the final dispersion was measured to be 1.9 A^{o} /mm.

The helium discharge tube was 1 1/8 ID with electrodes approximately 7" apart and filled with He at a pressure of 5 torr. A viewing port transverse to the discharge was provided with quartz windows 3/4" diameter. A compensating cell with the same size windows was placed in the opposing path of the interferometer.

The final arrangement is shown in figure 2. A power supply with a 15 KV 2 microsecond pulse excited the discharge tube. The dye laser was triggered either from the current pulse in the discharge or at a preset time relative to the discharge.

IV. RESULTS

The interferometer was offset from the white light fringe by approximately 1700 fringes producing an acceptable field with a fringe spacing of about 1 mm or 1.9 A^{o}. Under this condition the discharge

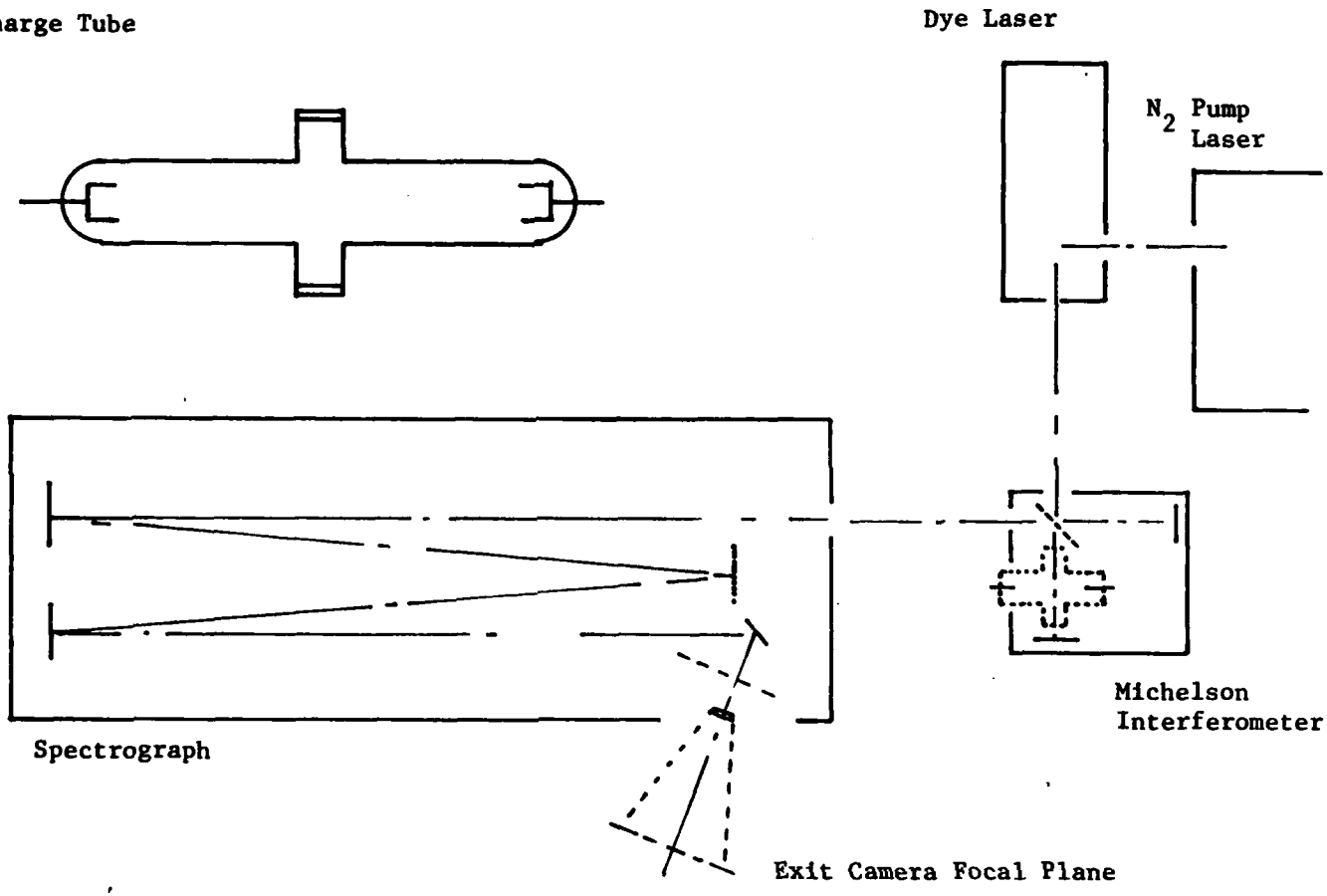


Fig. 2 Schematic of Hook Interferometer and Discharge Tube

was excited producing a peak current of approximately 500 amperes. The hook in the fringe pattern at the 5876 A^{o} wave length was then measured. From the tabulated values of f, (Wiese)[7] a value of 9.1×10^{12}/ cm^3 for the population density of the 2^3 P state was determined.

This value cannot be directly compared with the literature, but it is of the same magnitude as that reported by Miyazaki and Fukuda[5] for a discharge tube of helium at 10 torr and 12 KV, 800 amperes, 3 microsecond pulse. Changing the time of the dye laser firing by $\pm$ 1 microsecond relative to the peak current reduced the hook in the fringe pattern to nearly zero. More precise time measurements could not be made because of R. F. noise when the discharge tube and the nitrogen laser fired.

Replacing the Michelson with a Mach-Zehnder configuration has been completed, but time has not permitted the evaluation of it with this discharge tube.

V. CONCLUSION

Though most hook systems have employed a Mach-Zehnder interferometer, the Michelson type can be made to work satisfactorily. It should be pointed out that in the Michelson, the sampling beam passes twice through the discharge section, effectively doubling the path length.

VI. RECOMMENDATIONS

To fully exploit this technique, additional experiments are necessary to more precisely define the timing of the dye laser so that the time history of the excited states can be determined. To transfer the equipment to use the technique on a full scale discharge port will require considerable design effort.

Another area in which this research effort might go would be to

look at the effects of contaminants on population densities.

REFERENCES

1. D. Roschdestvensky, "Anomalous Dispersion in Sodium Vapor," Ann. Physik., Vol. 39, p. 307, 1912.

2. M. C. E. Huber, J. Opt. Soc. Am., Vol. 56, p. 1428, 1966.

3. W. C. Marlow, "Hakenmethode," Applied Optics, Vol. 6, No. 10, pp. 1715-24, 1967.

4. Darshan S. Dosanjih (Ed.), Modern Optical Methods in Gas Dynamic Research, New York, Plenum Press, 1971.

   See: D. Bershader, "Some Aspects of Refractive Behavior of Gases," pp. 65-83.

   J. H. Spurk, "Hook Interferometry Using a Single Plate Interferometer," pp. 113-116.

   M. C. E. Huber, "Interferometric Gas Diagnostics by the Hook Method," pp. 85-112.

5. K. Miyazaki and K. Fukuda, "Time-resolved Population Measurement with N_2 Laser-pumped Dye Lasers," Jap. J. Applied Phys., Vol. 10, pp. 1905-1910, 1977.

6. P. Van de Weijer and R. M. M. Cremers, "Experimental Determination of $6^3P - 6^3P$ Collisional Excitation Cross Sections for Line Emission in the Positive Column of D. C. Mercury Discharges," J. Appl. Phys., Vol. 53, No. 3, pp. 1401-1408.

7. W. L. Wiese and M. W. Smith and B. M. Glennon, Atomic Transition Probabilities, Vol. 1, Washington, D.C., U. S. Government Printing Office, 1966.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

LASER DAMAGE IN FILMS AND PLASTICS

Prepared by:	Dr. Thomas Wiggins
Academic Rank:	Professor
Department and University:	Department of Physics, Pennsylvania State University
Research Location:	Frank J. Seiler Research Laboratory
USAF Research Contact:	Lt Col Theodore T. Saito
Date:	August 26, 1982
Contract No.:	F49620-82-C-0035

Acknowledgements

The author wishes to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for the opportunity of spending the summer at the Frank J. Seiler Research Laboratory.

He also wishes to thank the personnel of that laboratory for their friendliness, cooperation and assistance. In particular, Dr. T.T. Saito, Dr. Albert Alexander and Dr. Kenneth Siegenthaler for their support and suggestions during the course of this work. Graduate student James Bogan, and undergraduates John Keating and J. Peter Holub also were of great assistance as was Earl Barr, Electronics Specialist. Jeffrey Hosken and Dr. Barry Crane were most helpful in the operation of the laser systems.

LASER DAMAGE IN FILMS AND PLASTICS

by

T. A. Wiggins

Abstract

Methods and techniques for a study of laser damage thresholds in thin films and plastics were investigated. To observe damage at 1.3μ m, the conversion of the 1.06μ m output of a Nd:Glass laser to the 1.3μ m wavelength by stimulated Raman scattering in NO was attempted. No output at 1.3μ m was detected probably due to the excessive line width of the laser. Techniques for the study of damage in plastics at 1.06μ m was prosecuted using a Nd:YAG laser. Commercially available types of clear plastics were tested including Lexan, Acrylic, and Uvex. A supposedly optical grade plastic was also tested. The highest damage threshold was found in Acrylic at 5×10^9 watts/cm^2. The damage in this and the other plastics appear to be due to cracks and burns at random positions in the laser beam showing small (<10 μm) defects to be the initiators of damage. A technique for producing multiple damage tests from a single beam was tested. Suggestions for continued work are included.

I. INTRODUCTION

The continually increasing requirements for energy and power to be delivered by laser systems calls for continually increasing damage levels for their components. In particular, coatings to enhance a mirror's reflectivity and to reduce the reflectivity of transmission elements appear to be the limiting factor in many systems. Thus, a study of thin film damage at several wavelengths of interest in high fluence systems has assumed a high priority. Measurements must be made at the design wavelength with carefully controlled conditions to provide meaningful data to guide the preparation of better coatings.

A second aspect is the weight and cost of optical elements. In particular, windows, prisms, beam-splitters, etc. could be prepared from plastics rather than glass or crystalline materials provided damage thresholds and other optical properties are commensurate with their intended use. Some success in the use of plastics has been reported.[1] It is important to verify these measurements and to conduct studies which may lead to even higher damage thresholds.

II. OBJECTIVES

The specific objectives of this study are two-fold: first, to explore the possibility of using stimulated Raman scattering in NO for the conversion of the output of a laser operating at 1.064 μm to a wavelength near 1.315 μm, the wavelength of the atomic iodine laser; and second, to explore the techniques of beam quality measurement and to use a well-quantified beam to study the damage threshold of plastic materials in the short-pulse, small-spot regime.

The motivation for these studies is: to provide a source of 1.315μm radiation of high intensity for testing of dielectric coatings; and, to determine the feasibility of the use of plastic elements for 1.06 μm laser systems.

III. EXPERIMENTAL

A. Laser Systems. Two systems were employed in this work. Both operate at 1.064 μm.

1. Nd:Glass. This high energy system manufactured by Spacerays was used in an attempt to produce Raman-shifted light at 1.315 μm. It is a water-cooled system with an amplifier stage that may be used. The oscillator is a two-flashlamp pumped rod 1x18 cm in size. A Brewster stack polarizer and a Q-switching Pockel cell are included in the cavity. It will produce up to five joules of energy with a pulse width of the order of 25 nanoseconds. The amplifier has a gain of ~10. It has an extremely large divergence of about 20 millirad.

2. Nd:YAG. This low energy system was constructed in this laboratory. A 5x45 mm flat-flat rod is air cooled and pumped with a single linear flash lamp. The resonant cavity of length 80 cm is formed with flat mirrors. A passive Q-switching dye is included in the cavity. The output can be effectively polarized by using three stacked glass plates at the Brewster angle. The unpolarized Q-switched output is about 30 milli joules in 15 nsec. The divergence is about one millirad.

B. Beam Diagnostics. Damage threshold is often reported in terms of the intensity at which failure is detected. Thus, measurement of energy, time, and the area into which the energy can be focused are important.

1. Energy and Time. These measurements are quite accurate and precise due to the availability of good calorimeters, photo-diodes and oscilloscopes. Generally, the accuracy in each is less than 15% with precision of 5% or less.

2. Measurement of the focussability of the beam presents a more difficult experimental task. Almost all methods assume that the beam has radial symmetry, a situation which strictly occurs only in principle. The minimum size of the beam (radius at which the intensity has decreased to $1/e^2$ of its maximum) is referred to as the waist. Thus, a measurement of the waist, ω_0 is the desired result. Several methods have been used:

a. Far-field divergence. If the cross-sectional area of the beam is measured, for example, by the size of the burn on a film, the waist can be inferred from $\omega_0 = \lambda z/\pi r$, where r is the radius of the beam at distance z from the waist. This is a very qualitative method.

b. Pin-hole scanning. Assuming rigorous radial symmetry, a measurement of the relative intensity as a function of radius at the position of the waist gives a direct measurement of the waist. This method can only be used with the best quality beams and has the disadvantage in requiring quite a few measurements in a beam of high reproductability.

c. Knife-edge scanning. Moving a knife edge across the beam gives similar results to that above. Radial symmetry must be assumed. An extension of this method has been used in the present experiments. If measurements are made at several positions with respect to the presumed position of the waist, the exact position as well as the size of the waist can be determined. If ω is the $1/e^2$ radius of a beam at distance Z, it can be shown[2] for Gaussian beams, that

$$\omega^2 = \omega_0^2 + (\frac{\lambda}{\pi\omega_0})^2(Z-Z_0)^2 \qquad (1)$$

Simultaneous solutions for several pairs of values of ω and Z allow a determination of Z_0 and ω_0 by requiring that the predicted value of ω_0 obtained from the intercept of a ω^2 versus $(Z-Z_0)^2$ plot gives the same results as solving for ω_0 from the slope of this plot.

d. A fourth method is under development. For beams of such wavelength as will affect photographic film, a photograph of the 90^o scattered light from a suspension of small particles should allow a measurement of ω_0. It has the advantage of requiring only one measurement. However, its feasibility has not yet been demonstrated.

C. Plastic Materials. Several readily available commercial plastics were tested to determine their damage threshold. These were Acrylic (Rohm & Haas), Lexan (Gen Elec), and Uvex (Eastman Kodak). In addition, two of supposed optical grade were also obtained. One was a plastic used as a ruby laser (694 nm) Q-switch and the other secured from Applied Products, Bristol, PA.

D. Technique for Multi-Spot Exposures. The determination of damage thresholds is a tedious process since many exposures of a material are required to select the intensity level at which damage does or does not occur. A special technique has been established which allows from 10-20 different sites on a sample to be exposed in a single shot. Of course, there is a loss of intensity but this may not be serious for some problems.

If a metal sheet containing a matrix of round holes is placed just before a focusing lens, the lens will form the Fourier transform of this array of holes which is itself an array of intense areas of spacing $\theta \sim m\lambda/d$, where m is an integer and d is the hole spacing. If each hole has a radius a, the observed intensity will smoothly decrease from unity at the center to zero at an angle $\theta = 0.61\,\lambda/a$. If the ratio is chosen to be about 5, a set of intense regions as shown in Figure 1 will be produced in the focal plane of the lens. The size of this array will be determined by the focal length of the lens. The angular size of each intense region is determined by the divergence of the incident beam as a lower limit and by the number of holes illuminated which sets an upper limit. In practice, if a region illuminated on the array covers a large number of holes, there will be only a small increase in the size of the intense areas compared to what it would be with no array of apertures present. The angular width due to the finiteness in the number of apertures is given by the spacing of the intense areas divided by the number of apertures in a straight line which are illuminated. Arrays with 60° and 90° orientations have been used.

The intensity of each spot can be calculated from

$$I = I_0 \left[\frac{2 J_1 (2\pi a \sin \Theta/\lambda)}{2\pi a \sin \Theta/\lambda} \cdot \frac{\sin (N\pi d \sin \Theta)}{N \sin (\pi d \sin \Theta)} \right]^2 \qquad (2)$$

where J_1 is a first order Bessel function and N is the number of holes in a straight line that are illuminated.

One limitation in the calculation of the intensities is the uniformity of the radii and spacings of the apartures. Uniformity of hole radii is the major problem. It could be removed if the holes are made of random size for which case the Bessel function will become a Gaussian distribution for a Poisson distribution of hole radii. That is, apodization is complete.

The reduction of intensity associated with the use of the array of holes can be visualized by noting that the first effect of the array is to reduce the total intensity by a ratio of $(\pi)(a/d)^2$ for a 90^o array and $(2\pi/\sqrt{3})\ (a/d)^2$ for a 60^o array. For d/a = 6, the transmission is about 10%. Second, the transmitted light is divided among some 15-20 spots, the center one receiving about 1/6th of the total. Thus, in one shot about 15 areas are illuminated with intensities that range from 1/50th of the available intensity to about zero. Due to redundancy, four or six areas will receive the same intensity.

An advantage to this technique in addition to providing multiple-intensity spots for damage testing is in providing multiple spots for assessing beam quality as determined by the adjustment of the laser's resonant cavity or the focus of a lens. Since one method of

measuring beam diameters is to "burn" a film or surface, if the beam's intensity is not properly adjusted, the burn area is either not present or enlarged in size. The present technique gives several intensities at one time.

Further, since damage often appears to be defect initiated, the rate of gathering data is enhanced since there is only a finite probability that an intense beam will coincide with a defect site. For example, if four spots receive the same intensity which is near the threshold for damage, only those which impinge on defects will produce damage.

Finally, since the separation on a surface of these intense areas can be selected by the proper choice of hole separation and focal length of the lens, the effects of conduction and diffusion can be studied since these physical processes will operate differently if several sources of heat or charge are provided compared with their operation from a single source.

IV. EXPERIMENTAL RESULTS

A. Stimulated Raman Scattering. The conversion of 1.064 μm radiation to 1.315 μm using stimulated Raman scattering in NO was not observed. Using the Space Rays laser beam of up to five joules in 25 nsec into an 80 cm cell at 400 mm pressure produced no observable radiation at 1.315 μm when: (1) a parallel beam was used; (2) the beam was focussed into the center of the tube with a 50 cm lens; or (3) when a plane mirror of low reflectivity at 1.31 and high reflectivity of 1.06 μm was placed at the output end of the tube and a reverse reflectivity mirror placed at the input end of the tube.

It is conjectured that the failure to see the converted radiation is due to the extreme line width of the laser system, the gain in a stimulated process being inversely proportional to the effective line width which here is dominated by the input line width rather than by the gas's Doppler or pressure width.[3]

A suggestion for a future experiment in the light of these results is included in Section V.

B. Damage in Plastics. Three different plastics were tested. Only for the acrylic was extensive data obtained since for the others, the damage threshold was lower by an order of magnitude and hence not interesting for present purposes. The source was the Nd:YAG focussed with a 15 cm lens.

In two of the plastics, stimulated Brillouin scattering (SBS) was observed.[4] These were the acrylic and ruby Q-switch plastics. These were the most brittle and hard and hence subject to the cracking needed to produce an intense acoustic wave. At threshold for damage, SBS was not observed. Measurements in these plastics showed less than 1% absorption for 3 mm thick samples at 1.06 μ m. The fourth sample remains to be tested.

Damage was observed in the following ways:

1. Observation of a bluish surface plasma during radiation when viewed through 1.064 μm laser safety glasses. This technique is essentially useless since the most intense plasmas appear to be produced by surface dirt. If this occurs, little radiation penetrates the sample.

2. A bluish plasma in the interior of a sample usually indicates that permanent damage will be found. This is very qualitative and subject to error.

3. Observation of enhanced 90^{o} scattering from a He-Ne alignment laser. This is also qualitative but if a scattering center is seen, permanent damage is almost always found.

4. Acoustic signal. Using a small acoustic transducer glued to the sample, a signal of the order of 100 times the noise level could be detected with intense damage. However, no sharp decrease in signal corresponding to a damage threshold could be observed. This is a similar result to that reported by Walker et al[5] for the scattering of the incident laser light or of an auxillary laser.

5. Scattering of incident radiation. This was detected but no sharp change in its intensity could be observed to act as a real-time indicator of damage.

6. Intensity of the transmitted beam or of an allignment laser. Both these techniques failed to indicate the presence of damage. If a damage center occurs outside of the most intense part of the beam, scattering will occur but the effect on transmission is small.

7. Microscopic examination. This is the most reliable and easily used for detection of damage, if damage is defined as a permanent change in the material. Observations at 100x were used to determine the power level for acrylic at which damage occurred. Using repeated shots, a power level of 1.4x10.5 watts, corresponding to an energy of 2.2 milli joules and a pulse duration of 15 nsec, damage

could be observed in about half of the shots. For twice this power, damage was always seen and for half this power, only three cases of single spot damage were observed in 14 trials.

This power level can be translated into an intensity with a knowledge of the waist size. Measurement as described in a previous section indicated a waist of 30 $\pm$ 10μ m. The intensity at the center of the beam's focus is thus 5 Gigawatts/cm^2, a result more than an order of magnitude lower than reported by Malenkov.[1] This smaller value may be attributed to dirt and defects in the sample.

Microscopic observations of the acrylic samples showed that cracks and fractures were produced at the highest intensity. Lower intensity produced voids and some burn areas. In the other plastics, the most prominent damage was black spherical regions presumably containing carbon.

Suggestions for procuring samples having higher thresholds are included in the next section.

C. Results using a hole grating. Figure 1 showed the intensity distribution produced by an array of holes whose separation to hole radius had a ratio of 5.2 for the 60° array and 3.8 for the 90° array. Photographs of these arrays produced by Buckbee-Mears Company in 18x21" sheets are shown in Figure 2. The photo-etching process produces amazingly regular holes considering the size of the arrays produced.

Damage observed using the Space-Rays system focussing with a 20 cm lens is shown in Figure 3. For the 60° array on a diamond turned nickel sample, several different sized damage sites can be seen. For the 90° array, damage on a diamond turned copper sample is shown.

Figure 4 illustrates the use of an array of holes in assessing beam quality. This damage region on copper clearly shows a mis-aligned laser cavity. Note that without the use of the array, no damage might have been seen for too weak a pulse, or a large burn might have been seen for too intense a pulse. The almost exact replication of damage implies that here the beam quality, not the quality of the surface of the sample, is the major factor in determining the damage morphology.

The ratio of the intensity in the central spot formed with the hole grating to that of the intensity of the focussed beam without the grating was measured using a pin-hole apertured photo detector and a He-Ne laser source. For the 60^{o} array the result was 1:65, and 1:31 for the 90^{o} array. These values are to be compared with the calculated values of 1:64 and 1:18 using the measured values of a/d. This agreement is considered to be satisfactory considering observed the variation in the radii, particularly for the 90^{o} array.

It is noted that a rigorous calculation of relative intensities produced by such a hole grating must take into account the variation of intensity across the input beam.

V. SUGGESTIONS FOR FURTHER WORK

Three areas of activity are suggested based on the summer's experience and progress:

a. Conversion of 1.06 μ m radiation to 1.315 μ m. Although effort to convert the radiation from the Nd:glass laser to 1.315 μm through stimulated Raman scattering was unproductive, one modification may be usable. If a low power Nd:YAG laser is modified using mirrors

having high reflectivity at 1.31 and low reflectivity at 1.06 μm, its output can be used as an oscillator and the Nd:glass used as a amplifier. It can have two stages and may produce adequate intensity for the proposed problem of damage testing of dielectric films for atomic iodine laser systems. As a minimum this technique can be used to measure the gain in Nd:glass at 1.31 μm.

b. Secure and test additional plastic samples at 1.06 μm with a well defined beam. These plastics may be commercially available or produced under controlled conditions to insure cleanliness. Some interesting parameters to explore would be their cross-linking, co-polymer composition, glass transition temperatures, brittleness, and the effect of additives. It is recommended that adequate characterization, including mechanical and optical properties, be obtained.

c. Preparation and use of hole gratings. The successful use of commercially available arrays of round holes with regular spacing in damage experiments and beam diagnostics indicates that even more precision can be obtained with careful preparation of two types of hole gratings of limited area. These are:

(1) Gratings with equal-sized holes. It was found that a limitation in the calculation of the relative intensities was the observed variation in hole diameters. The ones used were cut from an 18x21" sheet in which the regularity was surprisingly good considering the large area produced. Since only about an inch square is required, it probably can be produced by photoetching to a much greater precision. If this can be accomplished, the theory of the intensity distribution needs restudy to take account of the variable intensity across the array of holes.

(3) Gratings with random holes. It can be shown that if the radii of the holes follows a Poisson distribution, the grating will be apodized so that no zeroes of intensity due to an airy disc will be formed. This form will give an extremely large range of intensities for damage testing. Further, calculated relative intensities should be more accurate.

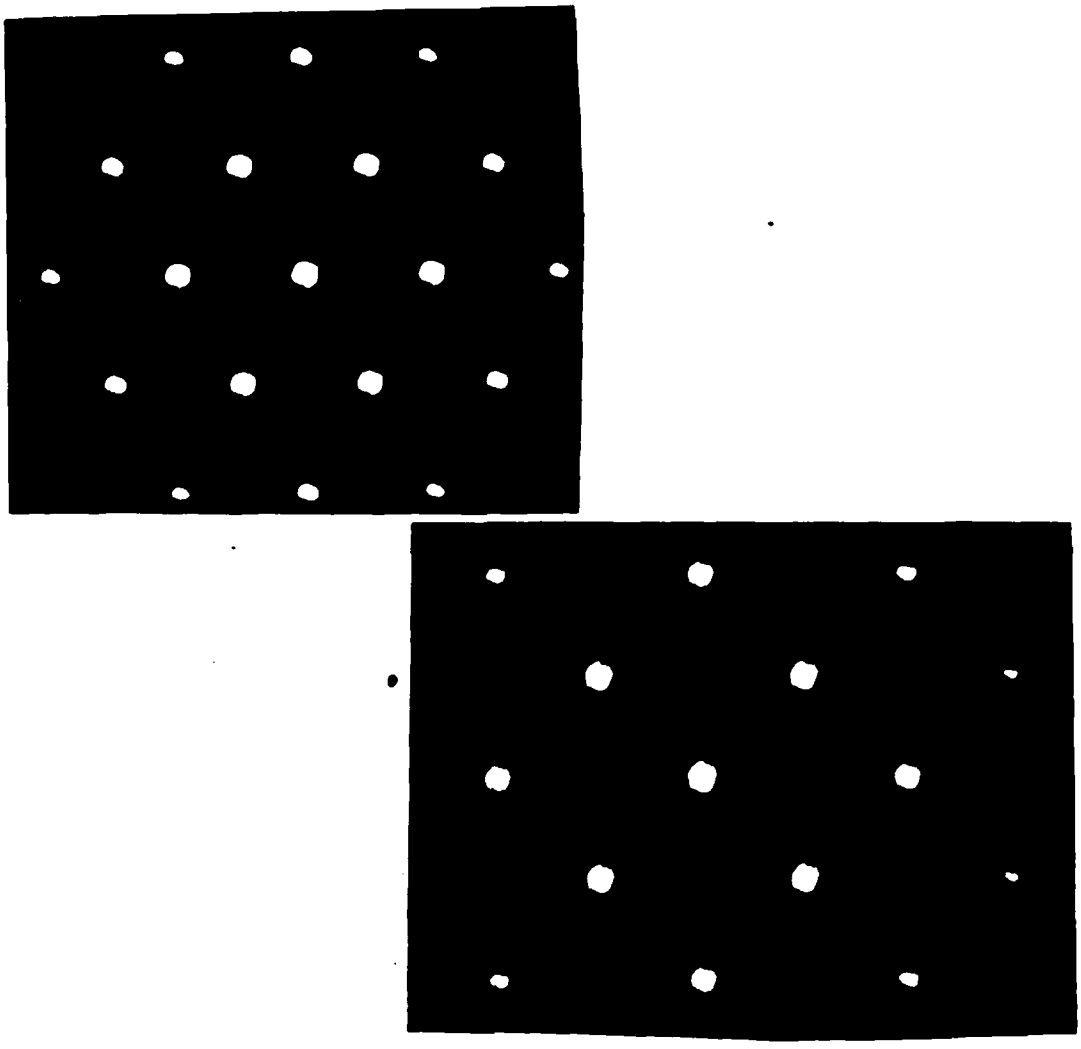


Figure 1. The observed intensity from 60° and 90° arrays of round holes

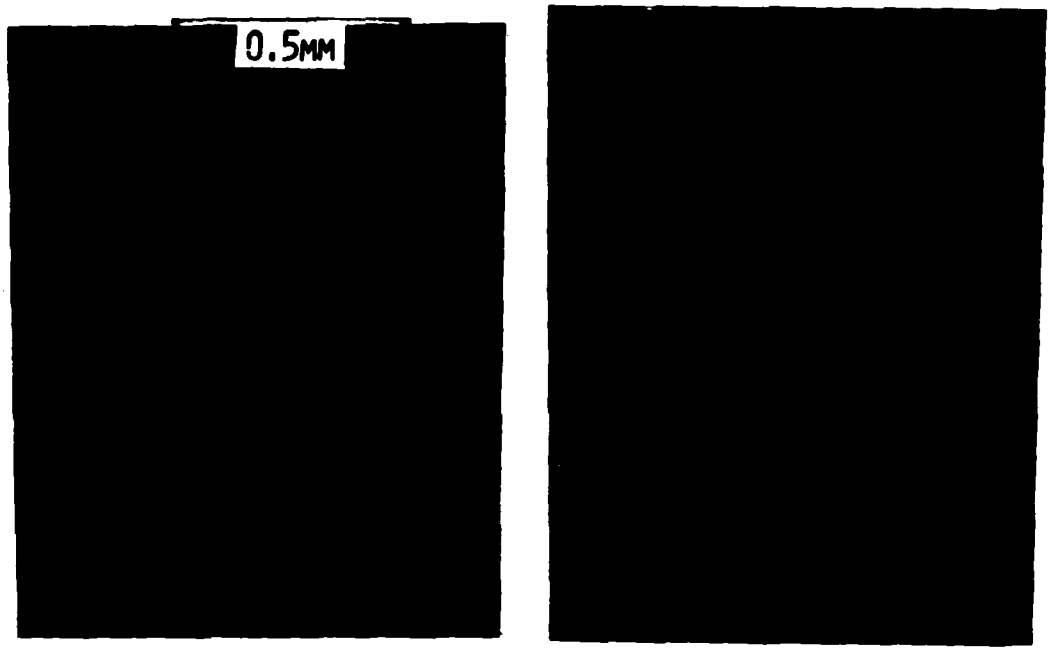


Figure 2. Appearance of the Arrays of Holes Used.
For the 60° array, d/a = $\frac{5.6\text{mm}}{1.1\text{mm}}$, and $\frac{4.0\text{mm}}{1.1\text{mm}}$ for the 90° array.



Figure 3. Damage Produced on a Nickel Sample Using the 60° Array, and on a Copper Sample with a 90° Array. Magnification is 63x.

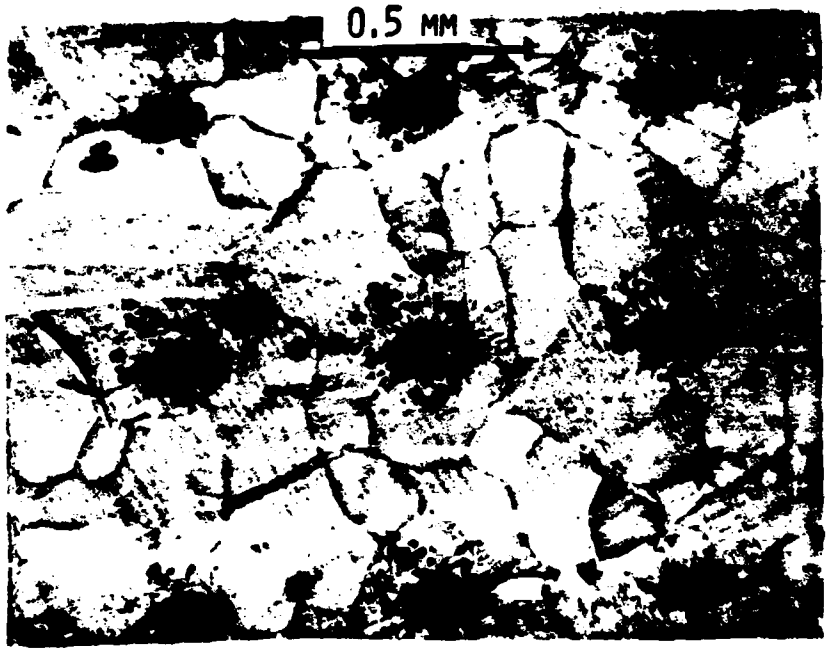
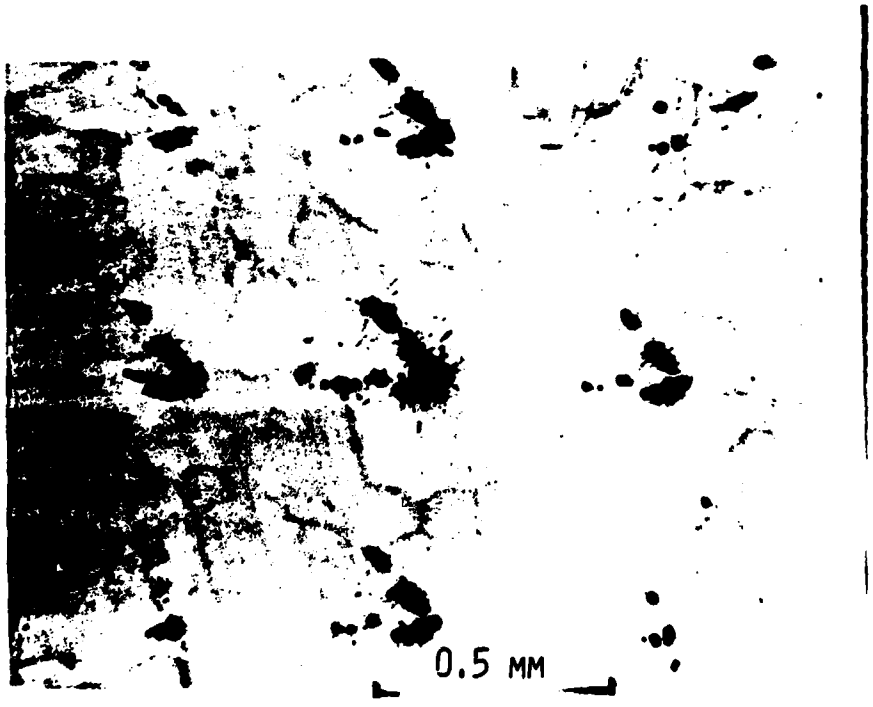


Figure 4. Damage to a Copper Surface Using a 60° Array When the Laser Cavity is Poorly Adjusted. Magnification is 63x.



References

1. K.M. Djumaev, A.A. Manenkov, A.P. Maslyukov, G.A. Matyushin, V.S. Nechitailo, and A.M. Prokhorov, "Transparent Polymers as New Class of Optical Materials for Lasers," given at Boulder Damage Symposium, 17-18 November 1981. To be published in Laser Induced Damage in Optical Materials 1981, NBS Special Publication.

2. A.E. Siegman, Introduction to Lasers and Masers (McGraw-Hill, New York, 1971) p. 309.

3. A.O. Creaser and R.M. Herman, Physics Rev Letts. 29, 147 (1972).

4. D.M. Smith and T.A. Wiggins, Applied Optics 11, 2680 (1972). D.R. Deitz and T.A. Wiggins, Applied Optics 43, 3631 (1971).

5. T.W. Walker, A.H. Guenther and P.E. Nielson, "Laser Induced Damage in Optical Materials: 1978", National Bureau of Standards Special Publication 541, A.J. Glass and A.H. Guenther, ed, p. 226.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

AN ALGORITHM AND ASSOCIATED DATA STRUCTURES FOR HIDDEN SURFACE ELIMINATION

Prepared by:	Dr. David H. Williams
Academic Rank:	Assistant Professor
Department and University:	Department of Electrical Engineering The University of Texas at El Paso
Research Location:	Air Force Weapons Laboratory
USAF Research Colleague:	Mr. Robert W. Conley Jr.
Date:	August 25, 1982
Contract No:	F49620-82-C-0035

AN ALGORITHM AND ASSOCIATED DATA STRUCTURES FOR HIDDEN SURFACE ELIMINATION

by

David H. Williams

ABSTRACT

A hidden surface algorthm is presented that uses a modified span coherence algorithm. The algorithm is structured such that simple processing is performed during the early stages so that the data is reduced during the later stages when more detailed processing is required. The hidden surface problem is addressed in an integral manner in which clipping, and the removal of backfacing surfaces are considered to be part of the hidden surface processing. The data structures are chosen to allow efficient processing at each stage, and to be modular so that new object attributes can be easily added. Linked lists are employed where dynamic memory allocation is required, and arrays or linked lists of arrays are employed where random access is important.

ACKNOWLEDGEMENTS

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research, and the Southeastern Center for Electrical Engineering Education, for a very interesting and productive summer at the Air Force Weapons Laboratory, Kirtland AFB, New Mexico. He would especially like to thank the personnel of ADPM at Kirtland for their assistance and hospitality; with special thanks to Mr. Robert W. Conley Jr. for his help and guidance, and many helpful discussions during this investigation.

I. INTRODUCTION:

Graphical display of three dimensional objects from a mathematical description has many useful applications in the simulation and modeling of scientific and engineering processes. In order for a graphical representation to be useful, however, hidden surfaces which are not visible to the human eye in a real scene typically should be removed in the graphical representation. This is necessary to improve clarity by removing the clutter of hidden surfaces, to remove the ambiguity which sometimes is added by the hidden surfaces, and of course to create a more realistic rendition of the desired representation.

Previous work has generated hidden surface algorithms which employ a variety [1,2,3] of approaches. Many, but not all of these methods assume that the image will be in raster format consisting of individual pixels (picture elements) formed into a rectangular array, and reproduced one line at a time. One common method requires the use of a depth buffer [1] in which each pixel of a line is generated by storing the intensity of the objects at that point back to front, with the front objects overwriting the intensities of the back objects. The end result is that only the intensities of the closest objects are stored and subsequently displayed.

Another common technique with several variations is that of span coherence [1,2,3]. Objects constructed of polygons are visible or invisible on a given scan line for the entire span that a given polygon is in front or in back of other polygons. Consequently, the test for visibility needs to be made only at the boundaries of each span, greatly reducing the amount of processing that is necessary for hidden surface elimination. Still other methods include the use of area coherence by Warnock [4], and the use of cluster coherence and separating planes by Schumacker et. al. [5].

In order for hidden surface elimination to be useful, a method must be chosen that is computationally efficient, is conservative of computer resources, and is generally applicable for a wide variety of imagery. There is no direct evidence that any of the given methods are superior for all applications. It is apparent, however, that the use of coherence greatly reduces the amount of processing that is required, particularly for complex scenes. Furthermore, the choice of data structures also has a direct bearing on processing efficiency, and should be chosen such that they are compatible with the hidden surface processing.

II. OBJECTIVES:

The purpose of this study was to design a hidden surface algorithm, with the appropriate data structures, which would be sufficiently general to display a wide variety of imagery, and which could be directly accessed by graphics users through subroutine calls, or through system supported utility programs. The primary research objective was the design of the algorithm and the associated data structures. A secondary objective was the implementation, which was not a direct requirement because of the time constraints of the 10 week research period. In fact, however, almost full implementation of the algorithm was obtained except for some final debugging.

III. APPROACH:

The general approach taken in the design was to configure the data such that it is processed efficiently at each step of the processing. Furthermore, in so much as is possible, the algorithm is structured such that simple processing is performed during the early steps when a large amount of data is present, so that the data is reduced as more detailed processing is required. An example of this is the processing of invisible backfacing surfaces, which on the average for closed polyhedrons, reduces the number of visible polygons for subsequent processing by one half, using only simple tests. The hidden surface problem is considered in an integral manner in which clipping and the removal of backfacing surfaces are considered to be part of the hidden surface processing. One result of this is that visibility information is passed from one step to the next, which (as just described) reduces the amount of data that must be processed during the actual hidden surface processing.

A modified span coherence algorithm is used for the actual hidden surface elimination, since this method is computationally efficient and has no inherent limitations which makes it unsuitable for a wide variety of applications. The computer language employed for the implementation has a direct bearing on the type of data structures which are suitable for a given application. Pascal was chosen for this application. Pascal supports dynamic memory allocation only through the use of pointers, therefore for information which required insertions and deletions, linked lists were employed. The use of linked lists also allows the data structures to be modular so that new object attributes such as color, shading

parameters, etc., can be easily added. Where random access of the data was the determining factor, arrays, or linked lists with elements composed of arrays were also used.

IV. DATA STRUCTURES:

The following is a description of the data structures, and also of the implementation which was generated to test the structural concepts. The description of a graphics scene is formed with objects, and two coordinate points, v the viewpoint, and c the center of interest, which define the line of sight of the viewer. Objects are constructed from one or more planar convex polygons which are described by an ordered array containing three to ten vertices. The objects describing a scene are joined together in a doubly linked list with each element as shown in figure 1. An object element is formed from a Pascal record which contains the pointers fwd and bak pointing to the succeeding and preceding object, respectively, the pointer poly which points to the first polygon record describing the object, and pointers prstrt and prfin that delimit the active polygon list which is used during the hidden surface processing. The object element also contains the coordinate points min and max which describe a bounding rectangle for the object, obnum which contains the object number, and npoly which contains the number of polygons which describe the object. Also included is ovflag which is a visibility flag that signals that the object is invisible if set to 1, or visible if set to any other value. A global pointer, obfirst points to the first element of the list which contains a NIL poly pointer. Each object element is added to this element so that a circular linked list is formed. Only the first element of the list has a NIL poly pointer.

The polygon information which describes each object is also formed into a doubly linked list which terminates with a NIL pointer in the forward direction, but which is circular in the reverse direction. The first element of the polygon list is pointed to by the poly pointer contained in the object element which is described by the polygon list. Each element of the polygon list is as shown in figure 2, and consists of a Pascal record which contains the pointers fwd and bak which point to the succeeding and preceding element of the list. Also included are nvtx and polyn which denote the number of vertices in the polygon, and the polygon

number, respectively. Pvflag is a visibility flag that has the same meaning and settings for the polygon that ovflag has for the object. Finally, vlist10 is an array that contains up to 10 vertex numbers which describe the polygon.

The polygon vertices are listed in vlist10 in a clockwise direction as they would be seen by a viewer. For example, the front vertices of a cube are listed in a clockwise direction, however, the rear vertices for the same viewpoint would be listed in a counterclockwise direction which would be clockwise if the viewpoint were to be moved to the rear. Since objects are typically formed from polygons which share common vertices, in order to share memory, the vertices are stored in a master vertex list, with the number associated with the vertex location stored in the polygon vlist10 array. Consequently, different polygons specified by the same vertex would have the same vertex number stored in their vlist10 array. The master vertex list consists of a singly linked list of arrays which can be allocated dynamically. Each of the arrays can store 1000 vertices and as a given array becomes full, the next is allocated and linked into the list. The global pointer, vfirst, points to the first array. An array was used for this data so that random access could be employed to obtain a given vertex without traversing through the other values.

As the processing proceeds, it is necessary to store boundary information for each polygon of each object. This information is also stored in the form of a doubly linked list, with each element as shown in figure 3. Included in each element are the current boundary x and z coordinates, xcoord and zcoord, and the increments, xinc and zinc which are added to the x and z coordinates as each scan line is traversed. Pointers o, p and obnd are stored which point to the object and polygon associated with each boundary, and to the other boundary of the same polygon. Fwd and bak links to the preceding and succeeding boundary elements are also stored. Also included are traverse, the traversal direction to the next boundary vertex; num, the vertex number of the bottom boundary vertex; nxty, the y value of the upper vertex, and val, the character l or r denoting whether the boundary is either left or right. For each scan line, the boundary list is employed to construct an active scan line list which specifies how each scan line is drawn. The scan line list is an array containing the necessary information for drawing each span

(section) of the scan line. Each element is a record which contains a pointer to the left x boundary of the span, a pointer pointing to the location containing the interior information of the span (color, intensity, etc.), and a pointer to the location specifying the right x boundary. These data structures specify how the data is processed and stored during the hidden surface elimination.

V. PROCESSING OVERVIEW:

The processing begins by specifying each object in an arbitrary world coordinate system. As each object is defined, a new object element for the object list is allocated, which in turn allocates new polygon elements for the polygon list associated with that object. The vertex numbers of each polygon vertex are stored in the polygon vertex array, with new vertices stored, and their associated vertex numbers created from the master vertex list as they are generated. The vertices in the master vertex list are then transformed into an eye coordinate system, in which the origin of the three dimensional coordinate system is translated to the viewpoint, and the coordinates are transformed from a right handed to a left handed coordinate system with z denoting depth along the line of sight. Back facing polygons are rendered invisible by setting the visibility flag, and the objects are clipped with respect to depth at the front and back clipping planes. The vertices are then transformed to add perspective, and clipped in the x (horizontal) and y (vertical) directions. Finally the vertices are transformed into device coordinates at which point the hidden surface processing is applied.

At this point the active object and polygon lists are created as subsets of the original object and polygon lists, that indicate which objects and polygons intersect a given scan line. The boundary list is generated, with each visible object and consequently each visible polygon associated with each object entering the processing (i.e. active list) as a scan line intersects its lower boundary, and leaving the processing as the scan line exceeds its upper boundary. The active scan line list is regenerated each time the active boundary list is changed, and each scan line is drawn from the information contained in the active scan line list.

VI. BACKFACING POLYGON ELIMINATION:

The elimination of backfacing polygons is considered to be part of the hidden surface processing. For scenes containing single objects the only hidden surface processing that is required is that necessary to eliminate the backfacing polygons that are hidden to the viewer. For scenes containing one or more objects occluded by other objects, the elimination of backfacing polygons eliminates approximately one half of the polygons which must be processed in succeeding steps.

The backface processing is as follows:

1. The first three vertices, vtx[1] - vtx[3], for each polygon of each object are accessed. (Each polygon contains at least three vertices which are assumed not to be colinear.)
2. Two vectors in the polygon plane are calculated:
   $w1 = vtx[3] - vtx[1]$ (1)
   $w2 = vtx[2] - vtx[1]$ (2)
3. The cross product is calculated to find the normal to the plane:
   $w3 = w1 \times w2$ (3)
4. The dot product is taken with the normal and a point on the polygon:
   $w3dotn = w3 \cdot vtx[1]$ (4)
5. If $w3dotn \leq 0$ the polygon is invisible to the viewer, and the polygon visibility flag pvflag is set to 1.
6. If the visibility flags for all of the polygons of an object are set, the object is invisible, and the object visibility flag, ovflag, is set to 1. (This situation would occur only for physically nonrealizable objects.)

VII. CLIPPING:

After the backface processing has been performed, the scene is clipped with respect to depth. Clipping in the horizontal and vertical direction is not performed at this time since with perspective, the clipping volume is a truncated pyramid [1]. Rather than performing the processing that compensates for this shape, the last part of the clipping is delayed until after the perspective transformation is performed, which changes the clipping volume into a rectangle. This avoids duplicate proces. .g. Clipping is performed one dimensionally in an identical manner for each of the three coordinates, and is as follows:

A. For each visible object, the maximum and minimum in the given dimension (calculated when the object was generated) are compared to the clipping boundaries to determine whether the object is completely visible, in which case the visibility flag, ovflag, is set to a unique negative value, or completely invisible, in which case ovflag is set to one. If the visibility flag has either of these values, or was rendered invisible by previous processing, the object list is traversed to obtain the next object for processing.

B. Otherwise, if the object is not completely visible or completely invisible, the maximum and minimum values of each polygon in the object are checked for complete visibility or complete invisibility.

1) If a polygon is completely visible or completely invisible, the polygon visibility flag, pvflag is set in an identical manner as for an object. If pvflag is either of the specified values, the polygon list is traversed to obtain the next polygon for processing.

2) Otherwise if it cannot be assertained that the polygon is completely visible or completely invisible the polygon undergoes the clipping process. For clipping, let wmin and wmax denote the clipping boundaries in the given direction, and let ic specify a clipping flag. The edges of the polygon are clipped in the order that their vertices appear in the polygon vertex list. Let A denote the first vertex of the edge and B denote the second vertex, with j denoting the dimension that is being clipped.

a) First ic is set to 0 (visible).

b) If A[j] and B[j] are both < wmin, or both > wmax, the edge is invisible and ic is set to 4.

c) Otherwise, if both A[j] and B[j] are $\geq$ wmin and both are $\leq$ wmax, the edge is competely visible, and ic is set to 0.

d) Otherwise, if B[j] < A[j], and if B[j] < wmin and A[j] $\geq$ wmin, then B is clipped to the point where the edge intersects wmin, and ic is set to 2 (trailing end point clipped.)
Or if B[j] $\leq$ wmax and A[j] > wmax, then A is clipped to the point where the edge intersects wmax, and ic is incremented by 1. Ic will then have the value of 3 if both endpoints have been clipped, or 1 if only the leading endpoint has been clipped.

e) Otherwise if A[j] $\leq$ B[j], if A[j] < wmin, and B[j] $\geq$ wmin, then A is clipped to the point where the edge intersects wmin, and ic is set to 1 (leading endpoint clipped).
Or if A[j] $\leq$ wmax and B[j] > wmax, B is clipped to the point where the edge intersects wmax, and ic is incremented by 2. Ic will then have the value 2 if only B is clipped, or 3 if both are clipped.

Clipping generates new vertices, consequently, as the edges of the polygon are traversed in order and clipped, a new vlist10 is constructed for each polygon. For this process, the running sum of ic, sumflag, is tallied.

3) If for a given edge ic = 0, the leading vertex (number) is stored in the new vertex list, if ic = 1, the new leading vertex is added to the master vertex list, and the vertex number is stored in the new list. If ic = 2 the leading vertex number is stored, the trailing vertex is stored in the master vertex list, and its vertex number stored in the new vertex list. If ic = 3, both vertices have been clipped, and both are stored in the master vertex list and their numbers stored in the new polygon vertex list.
4) After all of the edges of the polygon have been processed, if sumflag = 4*nvtx, pvflag is set to 1 (invisible). Otherwise, if sumflag is not 0, then the polygon vertex list and nvtx are updated to their new values.
5) All polygons of an object are processed in this manner. If all polygons are invisible, ovflag is set to 1.

VIII. HIDDEN SURFACE PROCESSING:

The actual hidden surface algorithm consists of a modified span coherence algorithm [1,2,3] which takes into account the coherence of objects in the horizontal direction. This processing is as follows:

A. The vertices in the master vertex list are transformed into real device coordinates and the maximums and minimums are recalculated for each visible object and each visible polygon.
B. Next, all visible objects are then sorted with respect to their minimum y value, using the value calculated in the previous step. A bucket sort should be employed for the sorting process.
C. All visible polygons of visible objects are then sorted with respect

their minimum y value, again using a bucket sort.

D. From scan line = ymin to ymax, the following processing is performed:

1) The active object list is first constructed and then modified as the scan lines traverse from ymin to ymax:

a) First, the starting pointer, orstart, is moved past invisible objects, and objects below the beginning or current scan line, to the first visible object intersecting or above the scan line.

b) Next the end pointer, orfin, is moved to point to the first visible object with a minimum y value > the current scan line. Orfin will then point to that value or to obfirst, the element with a NIL polygon pointer indicating the beginning and end of the list.

c) The global flag, DONE, is then set TRUE if orstrt = obfirst, indicating that all visible objects have been processed, or if orstrt↑.Y.min > ymax, indicating that no (more) objects will intersect any of the remaining scan lines.

d) The global flag, DRAWING, is set TRUE if orstrt↑.Y.min $\leq$ the scan line value, indicating that at least one object is intersecting the scan line and needs to be drawn.

e) Next, if DONE = FALSE, and DRAWING = TRUE the active polygon list is generated or modified for the given scan line. Otherwise the scan line is incremented to the next value, and control is transferred to step D. If the active polygon list is processed, for each visible object in the active object list:

i) If object↑.prstrt = NIL, then the starting and end pointers for the active polygon list, object↑.prstrt, and object↑.prfin are initialized as object↑.poly, the location of the first polygon in the object polygon list. Otherwise this step is skipped.

ii) Next, all boundaries from polygons ending below the current scan line (if any), are deleted from the active boundary list, and the boundary list change flag, CHNG is set TRUE.

iii) The starting pointer, object↑.prstrt is then moved past invisible polygons, and polygons ending below the current scan line to the first visible polygon intersecting or above the scan line.

iv) The end pointer, object↑.prfin is then moved to the first visible polygon with a minimum y greater than the current scan line. Object↑.prfin will then point to NIL (the end of the list), or to the first visible polygon above the current scan line.

v) The active polygon list is then traversed, and any polygons with a minimum y equal to the scan line value is inserted into the active boundary list, and CHNG is set TRUE. The left and right polygon boundaries are inserted as separate elements of the boundary list, and each of the boundary element values, xcoord, zcoord, etc., are initialized at this time.

vi) Next, the active boundary list is traversed, and the xcoord and zcoord values of each element are incremented by xinc and zinc respectively.

vii) The active boundary list is then sorted with respect to the x coordinate value, xcoord, contained in each element. A bucket sort to insert the elements, with a bubble sort to later resort the elements appears to be an efficient way to performing this processing. If one or more exchanges occur during the bubble sort, indicating an insertion. or crossing of boundary values from the previous scan line, CHNG is set TRUE. Otherwise, assuming no previous changes occurred, CHNG would be FALSE and no further changes to the boundary list (and subsequently to the active scan linelist) would be required. At this stage of the processing, the active object list points to objects which intersect the current scan line, sorted from minimum to maximum y, or the first object which will intersect a future scan line, or to NIL, indicating the end of the list. The active polygon list contains similar information about the polygons of each object sorted in the same manner, and the active boundary list contains the boundary information for each polygon which intersects the current scan line, sorted from minimum to maximum x.

viii) Next, if CHNG is TRUE, then the active scan line list is (re)generated. Each element of the active scan line list represents a span. A span is a unique nonzero length region of the scan line "spanned" by at least one polygon. The span elements are generated left to right from xmin to xmax, with pointers created which point to the left and right boundaries,

and to the visible polygon which spans the region, and is closest to the viewer. Consequently, each polygon which spans the region is searched to determine the one with the minimum value of z. This information is determined from the active boundary list.

ix) Finally, the active scan line list is traversed to draw the scan line on the display device.

IX. RECOMMENDATIONS:

The author feels that this algorithm can be efficiently applied to a wide variety of graphical images. A few comments are in order about its implementation, however. First, linked lists are appropriate where random insertions and deletions are made in a given data base, and double links assist in rearranging the data when sorting is necessary. It may be more efficient, however, to employ lists with single links, and to generate extra pointers for the sorting proeesses.

Secondly, insertion or bubble sorts were employed in the author's implementation, for expediency. However, a bucket sort should be employed for sorting the data in the y direction; and a bucket sort should be employed to generate the boundary list, and a bubble sort to modify it, to efficiently perform the x sort. If n is the number of elements to be sorted, a bucket sort is proportional to n [2] if the elements are originally randomly sorted, and both the bubble and bucket sorts are proportional to n if the data is nearly sorted. The bubble sort should require less operations for linked lists in the latter case, however, because less link operations should be necessary.

Two problems which have not been addressed by this algorithm appear suitable for future research. First, in some machines such as the Cray I, significant reduction in the processing time can be obtained by vectorizing the computations. Secondly, many physical objects are not completely opaque. It appears that this algorithm could be modified to include objects that are both opaque and transparent, greatly increasing the scope of its usefulness.

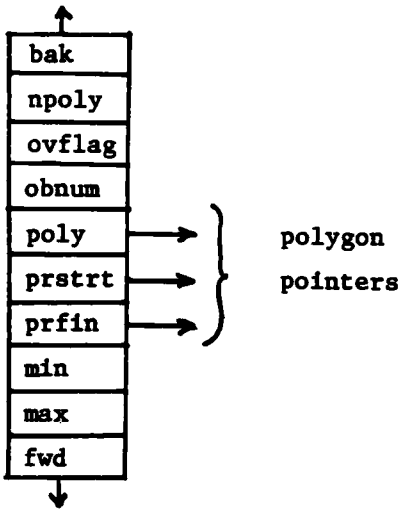


figure 1 Object Element

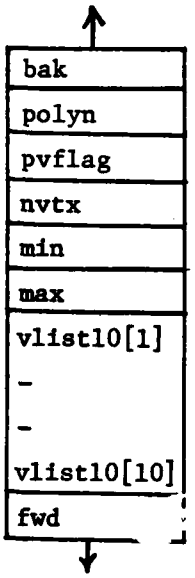


figure 2 Polygon Element

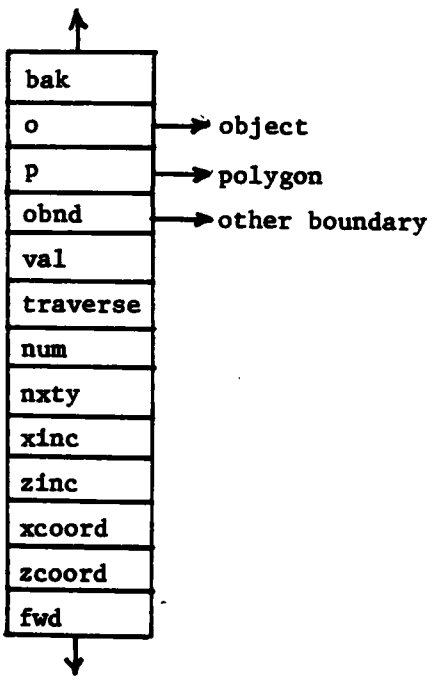


figure 3 Boundary Element

REFERENCES

1. Newman, William M., and Robert F. Sproull, Principles of Interactive Computer Graphics, McGraw-Hill, New York, N.Y., Second Edition, 1979.

2. Sutherland, Ivan E., Robert F. Sproull, and Robert A. Schumacker, "A Characterization of Ten Hidden Surface Algorithms", Computing Surveys, Vol. 6, No. 1, ACM, March, 1974.

3. Foley, James D., and Andries Van Dam, Fundamentals of Interactive Computer Graphics, Addison-Wesley, Menlo Park, California, 1982.

4. Warnock, J. E., A "Hidden-Surface Algorithm for Computer Generated Halftone Pictures", Computer Science Department, The University of Utah, TR 4-15, June, 1969.

5. Schumacker, R. A., B. Brand, M. Gilliland, and W. Sharp, "Study for Applying Computer-Generated Images to Visual Simulation", AFHRL TR-69-14, U. S. Air Force Human Resources Laboratory, September, 1969.

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

EFFORT TO PRODUCE METABORIC ACID VAPOR AND WATER VAPOR BROADENING OF CARBON MONOXIDE ABSORPTION LINES

Prepard by:	Dr. Robert E. Willis
Academic Rank:	Assistant Professor
Department and University:	Department of Physics Mercer University
Research Location:	Arnold Engineering Development Center, Engine Test Facility, Technology Department, Advanced Propulsion Diagnostics
USAF Research: Colleague:	Dr. C. C. Limbaugh and Mr. H. C. Walker, Jr.

Date:	September 8, 1982
Contract No.:	F49620-82-C-0035

EFFORT TO PRODUCE METABORIC ACID VAPOR AND WATER VAPOR BROADENING OF CARBON MONOXIDE ABSORPTION LINES

by

Robert E. Willis

ABSTRACT

Effort was made to produce HBO_2 in the gaseous phase by heating H_3BO_3 to over 1000 K in a vacuum. Attempts at detection were made with a Fourier transform infrared spectrometer. Based on what is known of its structure, a detailed prediction was made of the spectrum in the band centered at 2030 cm^{-1}.

The broadening by water vapor of the P(2), P(3), and P(7) lines of the fundamental vibrational transition of carbon monoxide was studied with a tunable diode laser infrared spectrometer at seven temperatures ranging from 400-1000 K. For each data point the collisional halfwidth was determined by fitting the line shape to a Voigt profile. The halfwidths were then fitted to $\gamma_L = \gamma_L^{\circ}(T_o/T)^n$.

Acknowledgement

I would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing me the opportunity to spend an interesting and fruitful summer at the Arnold Engineering Development Center, Arnold Air Force Station, Tennessee. I would like to acknowledge the center, especially the Technology Department, for its hospitality.

In particular I would like to thank Dr. C. C. Limbaugh for suggesting this area of research, and Dr. H. C. Walker, Jr. and Mr. Bill Phillips for their collaboration and guidance.

I. INTRODUCTION:

An area of current interest to the combustion technology community is the determination of exhaust gas species concentrations by means of non-interference techniques. One of the most promising techniques involves the use of high resolution tunable diode lasers in the accurate determination of absorption line strength and shape. The true absorption line shape is normally distorted by the spectroscopic instrumentation used to acquire the spectrum. However, the tunable diode laser spectromoter, with a resolution of approximately $10^{-4}cm^{-1}$, is able to produce an essentially undistorted spectra under conditions of temperature and pressure typical of a combustion environment.[1-3]

The shape of an absorption line of a particular chemical species depends on the frequency (wavenumber) of the line, the number of absorbing molecules in the path of the beam, the temperature of the abosrbing gas, and the identities and partial pressures of diluent gases.[4] Molecular collisions involving the absorbing molecules affect the line shape, and this effect differs according to the species of colliding molecule. Thus, a high resolution spectrum combined with the knowledge of the temperature and certain molecular parameters can provide the basis for the noninterference determination of species concentrations in the effluent streams of combusting systems.

A molecule possibly present in some combustion streams is metaboric acid, HBO_2, a species about which very little is known. The initial intent of this work was to use tunable diode lasers to make spectral measurements on HBO_2. This compound can be produced in the gaseous phase only by chemical reactions at high temperature. Since our equipment allowed only the measurement of temperature and total pressure, it would not have been possible to relate an experimentally determined line shape function to the partial pressures of absorbing and foreign gases. But it was hoped that the approximate magnitude and dependence on temperature of the halfwidth of HBO_2 lines could be determined. As little is known about the structure of the HBO_2 molecule, molecular parameters such

as bond lengths and angles could be determined from frequency measurements of a number of spectral lines.

Since the tunable diode laser can only be swept very slowly over an extremely small spectral region, it is not a good instrument with which to search for lines where frequency isknown only approximately. Therefore, a Fourier transform infrared spectrometer was used to search for the presence of HBO_2 absorption.

The efforts to produce HBO_2, which proved futile, are documented in Section IV of this report. Section III details a prediction of the HBO_2 spectra based on what is known of its structure.

After six weeks of effort to make HBO_2, it was decided to switch to a project that was fairly certain to generate data in a short period of time. This work involved the analysis of carbon monoxide absorption lines broadened by water vapor. The P(2), P(3), and P(7) lines of the V = 0 - 1 vibrational transition of CO (2111 cm^{-1} band) were each analyzed at seven different temperatures. Each absorption line was fitted to a Voigt profile [5] by a least squares procedure that yielded the Lorentzian (collisional) halfwidth, denoted γ_L. The Lorentz halfwidths were then fitted to

$$\gamma_L(T) = \gamma_L^{\circ}(T) \ (T_o/T)^n \qquad (1)$$

an approximate temperature dependence derived from the Anderson-Tsao-Curnutte theory by Birnbaum[6], where n is the temperature dependence exponent and T_o the reference temperature. This form has been assumed by several authors for the temperature dependence fo CO line width.[1-3,7-10]

II. OBJECTIVES:

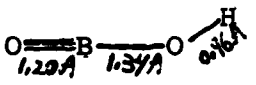
There were three main objectives to the initial project. The first was to make a prediction of the HBO_2 spectra of the 2030 cm^{-1} band. The second was to perform line broadening measurements on a number of HBO_2 absorption lines in the 2100 cm^{-1} region. From such measurements the line strength and Lorentzian halfwidth could be calculated. The halfwidth could be related

to the temperature of the absorbing gas, the total pressure, and the frequency of the line. The third objective was to measure the frequency of a number of HBO_2 lines in the 2100 cm^{-1} region and from this data calculate values for the bond lengths and bond angles of the HBO_2 molecule.

The second project had as its primary objective the determination of the Lorentzian halfwidth of three CO lines broadened by water vapor. The P(2), P(3) and P(7) lines of the fundamental IR band of CO were each studied at seven temperatures ranging from 300-1000 K. After adjustment of the average halfwidth at each temperature to its value at atmospheric pressure, formula (1) was used to determine the temperature dependence of γ_L and the halfwidth at standard temperature and pressure, γ_L^o.

III. PREDICTION OF THE EXPECTED HBO_2 SPECTRUM:

The infrared spectra of HBO_2 was predicted based upon what is known about its structure. The structure of the HBO_2 molecule has been debated in the literature.[12] The most convincing argument is that presented by White, et al, [11] who suggest the structure pictured below.



The infrared band at about 2030 cm^{-1} is excited by the fundamental vibrations of the double bond when the boron nucleus is the dominant isotope species B^{11}. About 20% of naturally occurring boron is B^{10}, for which the O===B stretch is at about 2080 cm^{-1}.

From the structure shown above the spectral constants of $HB^{11}O_2$ were claculated to be A = 26.346 cm^{-1}, B = 0.3078 cm^{-1}, C = 0.3043 cm^{-1}. The structure is thus that of a very prolate, slightly asymmetric rotor. The value of Wang's asymmetry paramater is -0.99970. This parameter, defined by $\kappa = (2B-A-c)/(A-C)$, has the value of -1 for a prolate top and 0 for the most asymmetric rotor. To a first approximation the infrared spectrum of a rotor where asymmetry is so slight is the same as that of a symmetric top.[18] With the molecule viewed as a symmetric top, the axis of symmetry is inclined to the O-B-O line by an angle of only 1.53°. Because the 2030 cm^{-1} band is excited by a change in the dipole

moment along the O-B-O line, which is nearly parallel to the symmetry axis ($\Delta K = 0$) will be much stronger than perpendicular transitions,[18] which were assumed to be unobservable in this band. If HBO_2 were a rigid symmetric top, the P and R branches of its spectra would consist of a series of equally spaced lines separated by $2B = 0.616\ cm^{-1}$. All lines of the Q branch would be degenerate, producing one strong line at the center of the band. Because $I_A/I_B = 0.009$, the intensity of Q branch would be about 0.010 that of the P or R branches[18], but since the lines would be unresolved, this might still be the strongestfeature of the spectrum. From a study of the partition function at 1000K, it was found that the peak population would occur at J = 33, and levels as high as J = 60 would have significant populations. This simple picture is complicated by three factors: the vibration-rotation interaction, l-type doubling, and the slight asymmetry of the molecule. These factors are discussed in order below.

Due to the vibration-rotation interactions and the anharmonicity of the vibration, lines with the same value of J and ΔJ are split according to the value of $|K|$. Herzberg[18] gives the following formulas:

$$\nu_R = \nu_O + \left[(A_v' - A_v'') - (B_v' - B_v'')\right] K^2 + 2B' + (3B' - B'')J + (B' - B'')J^2 \quad (2)$$

$$\nu_P = \nu_O + \left[(A_v' - A_v'') - (B_v' - B_v'')\right] K^2 - (B' + B'')J + (B' - B'')J^2 \quad (3)$$

where ν_R is the frequency of R branch lines, ν_P is the frequency of P branch lines, the single prime applies to the upper state and the double prime to the lower state. For parallel transitions A' should equal A", but B' will differ from B". For a given value of J, the levels with higher K have a lower population. The fraction of molecules in a particular level is given by

$$F_{J,K} = \frac{F_v(2J + 1)2 \exp(-E_{J,K}/KT)}{Q_r} \quad (4)$$

$$E_{J,K} = h\left[BJ(J + 1) + (A-B)K^2\right] \quad (5)$$

where F_v is the fraction in a particular vibrational state and Q_r is the rotational partition function, which for HBO_2 has the value 2.06×10^4. Equation (4) shows that even at 1000K only low values of K will have observable populations. If $B'-B'' = 10^{-3}cm^{-1}$, a typical value for similar molecules[11], lines with K = 10 would be shifted by 0.10 cm^{-1} from those with K = 0. At pressures common to a combustion environment the only observable effect of the vibration-rotation interaction would be a broadening of each spectral line. The broadening would be less the lower value of J.

Another complication arises from the phenomena of l-type doubling, an effect arising from degenerate vibrations in symmetric tops. Because of Coriolis interactions between vibrations and rotations, a splitting of levels with same $|K|$ occurs which is known as l-type doubling. HBO_2 is so nearly symmetric that its bending modes would be close to degenerate. The general treatment of l-type doubling is quite complicated, but the greatest splitting occurs for $K = \pm 1$, for which a simple approximate expression can be given:[17]

$$\Delta\nu = 2q(J + 1) \qquad (6)$$

where $q = 2B^2/\omega$. Taking ω to be 700 cm^{-1}, reported by White, et al, for the O-B-O bend, the value of q is $2.7 \times 10^{-4}cm^{-1}$. The splitting for other values of K is much less. The l-type doubling would be negligible for low values of J, but might result in a significant broadening for higher values of J.

Finally, we consider the effect of the slight asymmetry. For the case of very slight asymmetry the only change from a symmetric top resolvable at combustion pressures is a splitting of the $|K| = 1$ lines. This is similar to the effect of l-type doubling except that the dependence of the splitting on the magnitude of K is much more pronounced. An approximate general expression for the asymmetry splitting of the K levels has been given by Wang[17]

$$\Delta W(b) = \frac{b^K (J + K)!}{8^{K-1} (J-K)! \left[(K-1)!\right]^2} \left[A - \tfrac{1}{2}(B + C)\right] \qquad (7)$$

where for prolate tops

$$b = (C-B)/(2A-B-C) = (K+ 1)/(K- 3) \quad (8)$$

and K is K_{-1} for a near prolate top . Equation (7) reveals that the degenerate level splitting decreases with increasing K. For HBO_2 the value of b is 6.67×10^{-5}. As an example of the dependence on K consider the case J = 30, K = 1 and 2. For K = 1 the two lines are separated by 1.6339 cm^{-1}, while for K = 2 they are separated by only 0.0126 cm^{-1}.

In summary, the strongest lines of the 2030 cm^{-1} band of HBO_2 will consist of P and R branch symmetric top lines separated by 0.616 cm^{-1}. As discussed above, these could appear broadened by the vibration-rotation interactions and l-type doubling. The Q branch will be effectively degenerate, and probably no stronger that the stronger individual P and R branch lines. Due to the slight asymmetry the $|K| = 1$ lines will be widely split.

IV. EFFORT TO PRODUCE METABORIC ACID VAPOR:

A thorough search through the literature revealed that very little is known about the production of metaboric acid, HBO_2, in the gaseous phase, especially under experimental conditions available at this laboratory.[11,12,13,14,15] The method most pertinent to our experimental arrangement involved the heating of B_2O_3 to 1250-1460 K in a non-reactive cell followed by the addition of dilute amounts of water vapor.[11,12] When the readily available compound orthoboric acid, HBO_3, is heated slowly it loses water to form HBO_2, which may exist in one of three crystal modifications.[16] Dehydration continues as the temperature rises above 150°C, yielding viscous liquid phases believed to consist mostly of B_2O_3. Since with our experimental arrangement, introduction of water vapor would expose a cold surface where

B_2O_3 vapor would condense, we attempted to produce HBO_2 by heating H_3BO_3 in a vacuum.

A Fourier transform spectrometer, which can rapidly scan over a wide region, was used to search for the HBO_2 lines. The band centered at 2030 cm^{-1} is accessible with both the Fourier transform and tunable diode laser spectrometers. Renovation of a recently acquired instrument was completed during my first week, and much of the effort on HBO_2 involved simply learning to use this instrument. The sample cell was the same as that described in Section V.

The spectrometer records the intensity of radiation as a function of path length. Infrared radiation from a blackbody source is passed through the heated sample cell. The intensity is sampled at 8192 points as the path length is varied by approximately 4 cm. The data is digitized and recorded on magnetic tape. The analysis, performed by a computer, involves taking the Fourier transform of the raw data, which yields detected intensity as a function of frequency.

Listed below are a number of changes made to improve the performance of the spectrometer. Nitrogen purged pipes were added to the optical path to minimize the absorption by atmospheric CO_2, the spectrometer mounted on a plexiglass platform to ease alignment, and mounted on shock absorbers to reduce the effect of floor vibrations caused by large compressors elsewhere in the same building. The apertures were reduced from 1 in. diameter to ½ in. because it seemed that the detector was seeing part of the hot cell wall. A water-cooled shutter was made to shield the blackbody from the detector when the background spectra was recorded. Finally, the elements in the furnace were rewired to allow higher temperatures to be reached.

Each attempt to detect the production of HBO_2 proceeded as follows. Crystals of H_3BO_3 were put in a small sealed tube, holes drilled in one side, and the tube placed in the sample cell. As the temperature was raised, water vapor, and later B_2O_3 vapor, should have escaped through the holes. Multiple scans were taken

with the cell cold and at various high temperatures, the highest being 1070K. The computer would average these scans. At the higher temperatures, scans were taken with the 600K blackbody viewed through the cell and with the blackbody shuttered. Subtracting the latter from the former gave the intensity of the transmitted radiation as a function frequency. Comparison of the cold and hot transmission spectra showed a general attenuation from 2100-1800 xm^{-1}, the high frequency range of the detector response. The explanation of this is that the transmission characteristics of saphire change with temperature, so that when the cell was hot the windows were absorbing some fo the blackbody radiation above 2100 cm^{-1}. Upon opening the cell, it was found that the copper O-rings had reacted to form the reddish compound Cu_2O.

V. CO LINES BROADENED BY WATER VAPOR:

The intensity of radiation that has passed through an optical path length l of a homogeneous layer of sample is given by the Lambert-Beers law

$$I(\nu) = I_o(\nu)e^{-k(\nu)l} \tag{9}$$

where $I_o(\nu)$ is the radiation intensity incident on the sample, $I(\nu)$ is the intensity transmitted by the sample. The absorption coefficient, $K(\nu)$, also called the line shape function, is a parameterized function of the wavenumber. The key parameter of this function, the halfwidth, is defined as half of the width at half-height of the absorption coefficient. In the limiting case of no collisional broadening (low density, high temperature), the line shape is given by the Doppler function [17], while in the limiting case of collisional dominance, the line shape is described by the Lorentzian function.[17] A general form known as the Voigt line shape includes both regimes:[5]

$$k(\nu) = \frac{S[\ln 2/\pi]^{1/2}}{\gamma_D}\frac{a}{\pi}\int_{-\infty}^{\infty}\frac{\exp(-y^2)\,dy}{a^2+(x-y)^2}, \tag{10}$$

where γ_D is the Doppler halfwidth

$$\gamma_D = (2kT\ln 2/M_a c^2)^{1/2}\nu_o,$$

M_a is the mass of the absorbing molecule, γ_L the Lorentz halfwidth, $_o$ the frequency of the line center,

$$a = \ln 2^{\frac{1}{2}} (\gamma_L / \gamma_D), \text{ and} \tag{12}$$

$$x = [\ln 2]^{\frac{1}{2}} (\nu - \nu_o) / \gamma_D. \tag{13}$$

The line strenth, S, for an isolated absorption line, is defined by the relation[17]

$$S = \int_{-\infty}^{\infty} k(\ ) d\nu \quad . \tag{14}$$

The effect of collision broadening on the line shape can be characterized by a Lorentz broadening parameter $_L(cm^{-1}atm^{-1})$. This parameter is defined as the collisional halfwidth attributable to the gas at 1 atm. pressure. $_L^o$ can refer to self-broadening or broadening due to a foreign gas. In the present study γ_L^o was determined for the water vapor broadening of carbon monoxide lines. The net halfwidth of a mixture of gases can be calculated by means of the expression

$$\gamma_L = \sum_i \gamma_{L_i}^o P_i \tag{15}$$

where P_i is the prtial pressure of the i^{th} component of the mixture.

A theoretical halfwidth can be calculated by means of the theory of Anderson[19] as expanded by Tsao and Curnutte.[4] From this theory Birnbaum derived the temperature dependence equation (1), to which the data were fitted.

The spectra were taken by a Laser Analytics LS-3 high resolution tunable diode laser spectrometer, which has a $10^{-4}cm^{-1}$ resolution. The sample cell consisted of saphire windows bonded to Kovar flanges which were welded to a 3 in. long stainless steel pipe. The temperature is monitored by a thermocouple inside the cell. A schematic diagram of the optical system is shown in Fig. 1. The lasers could be tuned over a range of 50 cm^{-1} by adjusting the temperature of the cold head in which the diodes were contained. A particular laser mode could be tuned over a range of about 1 cm^{-1} by modulating the diode injection current.

After passing through the sample cell the beam passed through a calibration cell that could be filled with CO to determine the detected power level at saturation. A 3 in. germanium etalon mounted on a slide could be introduced into the beam. The etalon produced fringes with a spacing of 0.0159 cm^{-1} which can be used to related diode injection current to output wavenumber.

Since the amount of CO admitted to the sample cell could not be measured, the line strength could not be determined. Water vapor was obtained by heating liquid water to the temperature corresponding to the desired vapor pressure. The vapor was supplied to the sample cell through previously evacuated heated transfer lines and valves which were maintained at 400K to prevent recondensation. The pressure was also monitored by means of a heated pressure transducer. The water vapor apparently condensed somewhere in the supply line, as the pressure steadily dropped while the spectrometer swept through the line.

The data acquisition system consisted of an analog signal processor, a minicomputer, and a digital tape drive to record the data on 7-track tape. Following each data point taken at known pressure and temperature, a 100 percent transmission trace was taken. Fig. 2 shows an example of raw data, including the 100% transmission and etalon traces.

The amount of radiation absorbed by the gas sample is the difference between the amount of radiation transmitted to the detector with and without the gas sample present in the beam path. First the 100% transmission trace was smoothed using a least-squares procedure, then the absorption line fitted to a Voigt profile by a least-squares procedure that used as many as 75 iterations. The line strength and Lorentzian halfwidths were the adjustable parameters whose values were determined by the fit. Fig. 3 shows the same data point after analysis.

Data for each line were taken at 100K intervals from 400K to 1000K. The values of $\gamma_L^{\circ}(T)$ were then fit to the (1) by a procedure that yielded the temperature coefficient n as the adjustable parameter. The results are summarized in Table 1,

where the standard deviation refers to the fit to (6). $_L(T)$ for each line is plotted as a function of T in Figures 4,5 and 6. Error bars for each point represent one standard deviation for the entire fit.

VI. RECOMMENDATIONS:

To make HBO_2, one probably needs a temperature higher than 1070K. Any cool surface on which B_2O_3 would condense is certainly to be avoided. Unless HBO_2 were produced in large quantities, it might not be detected with the Fourier transform spectrometer. Individual lines, which are so numerous, may not be resolved, and a general absorption over the band would be difficult to distinguish from attenuating by the saphire windows. But in large enough quantities the HBO_2 band, centered at 2030 cm^{-1}, would be distinct from the saphire attenuation, which increases with frequency.

Water vapor broadening of CO lines could be investigated experimentally for higher rotational states to determine if these is a dependence on J of the halfwidths. The accuracy of the data could be improved by placing the pressure transducer closer to the hot cell and eliminating any spots where water vapor might condense. It would be interesting to compare the experimental data to a theoretical calculations of the Lorentzian halfwidths based on the Anderson-Tsao-Curnutte theory.

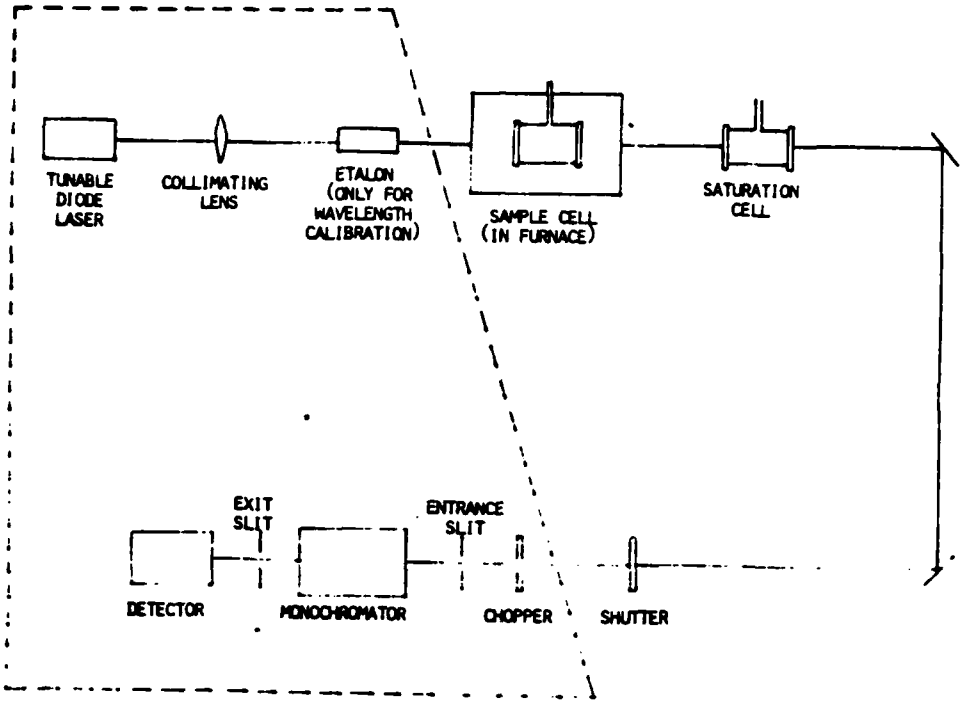


FIGURE 1 - SCHEMATIC OF LABORATORY OPTICAL SYSTEM

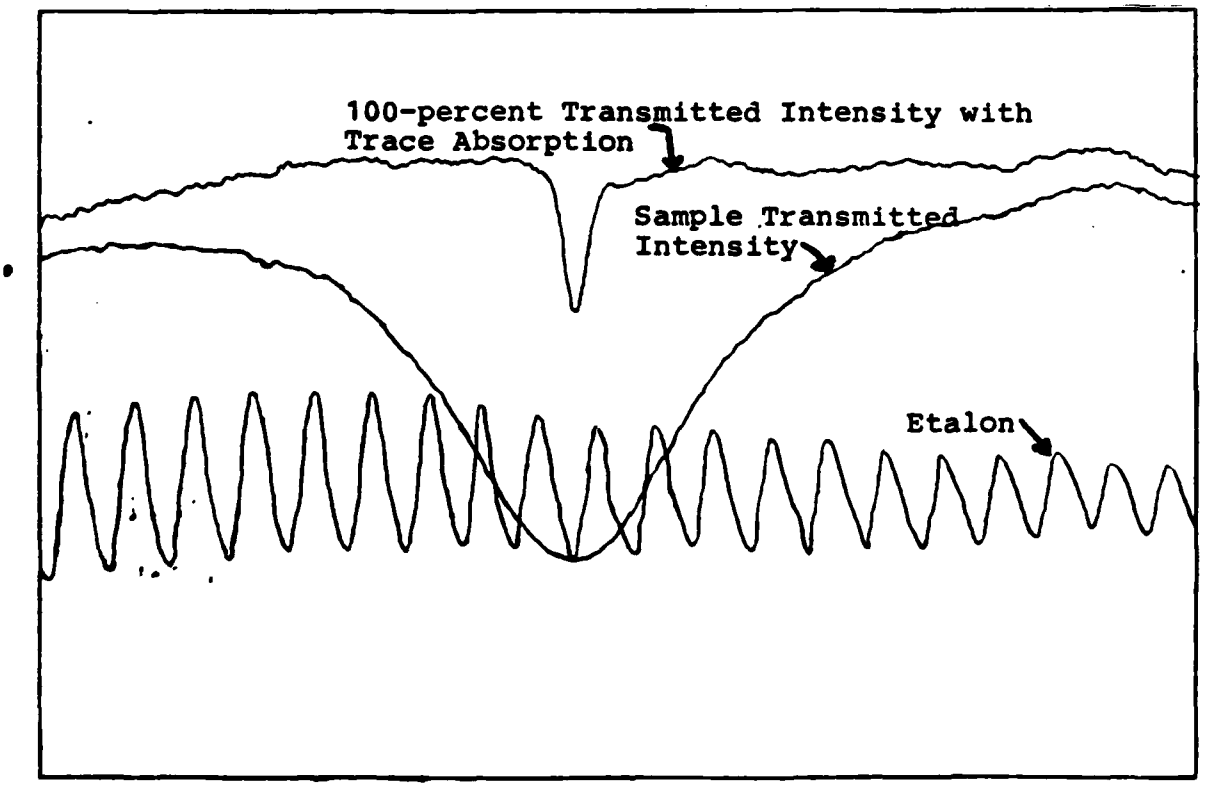


Figure 2. An example of raw data.

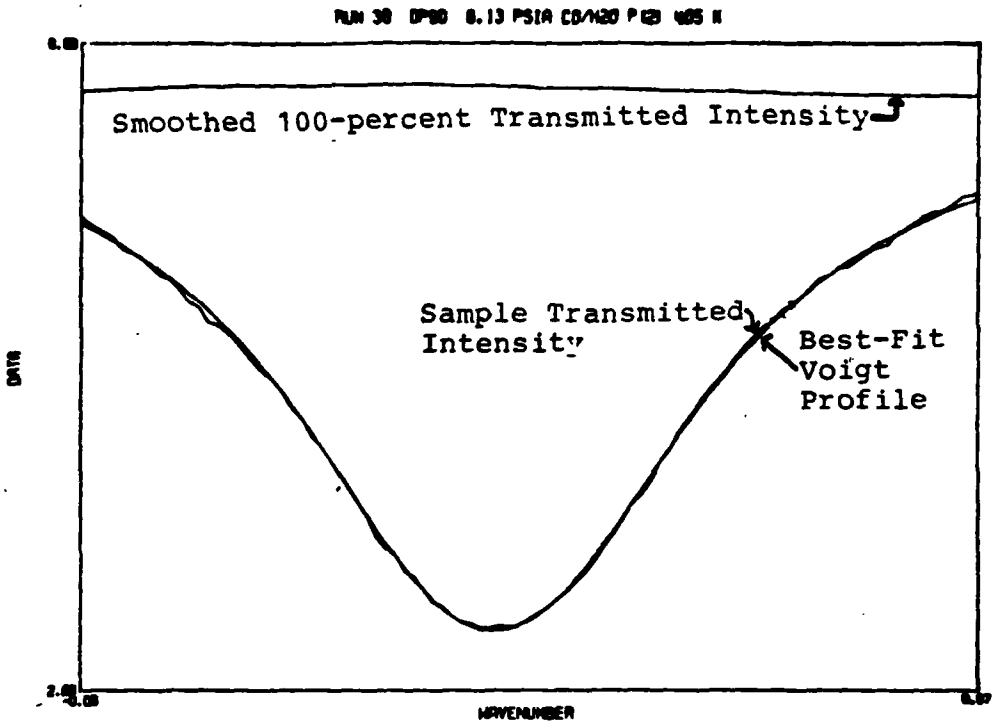


Figure 3. An example of the reduced data.

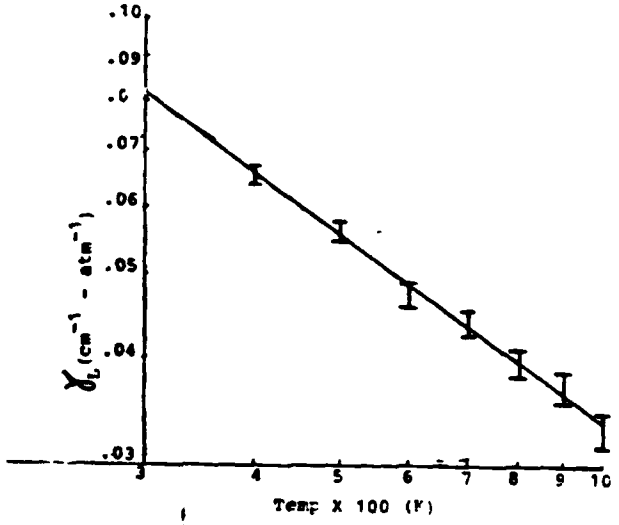


Figure 4. Temperature dependence of the experimental half-width for the P(2) CO line broadened by water vapor.

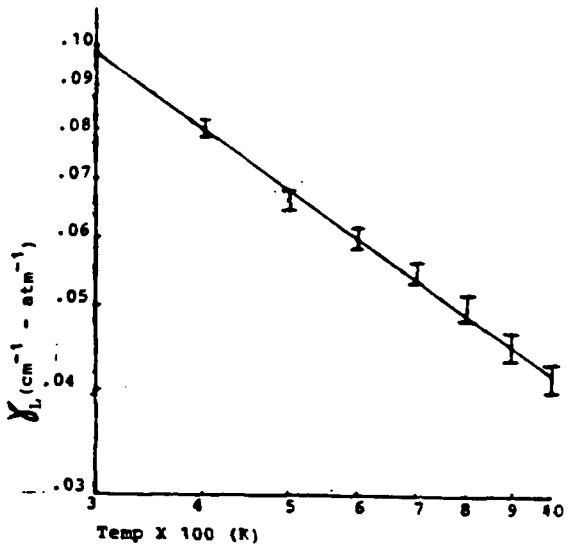


Figure 5. Temperature dependence of the experimental half-widths for the P(3) CO line broadened by water vapor.

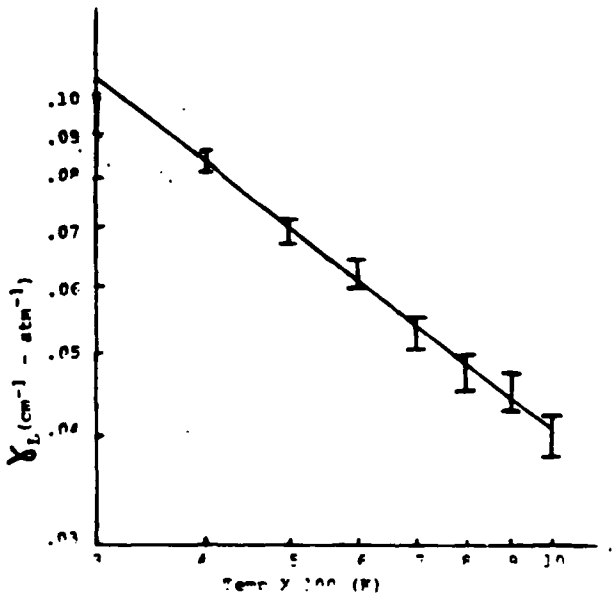


Figure 6. Temperature dependence of the experimental half-widths for the P(7) CO line broadened by water vapor.

Table 1. Summary of Experimental Water Vapor Broadened Halfwidths

J	T (K)	γ^{exp}_{ave} X10^3(cm^{-1}atm^{-1})	Standard Deviation	# Data Points	n
	400	65.54	0.84	7	
	500	56.38	1.19	7	
	600	47.41	1.42	7	
2	700	44.94	1.02	7	-0.7342
	800	39.54	1.34	7	
	900	37.01	1.44	7	
	1000	33.05	0.70	7	
	400	79.91	2.74	12	
	500	66.10	0.80	5	
	600	59.74	1.18	10	
3	700	54.26	1.00	6	-0.7017
	800	49.68	0.69	7	
	900	45.03	1.29	8	
	1000	41.36	1.16	9	
	400	83.71	2.53	11	
	500	69.41	2.83	6	
	600	62.09	0.94	7	
7	700	53.03	2.21	7	-0.7813
	800	47.77	1.72	6	
	900	45.46	1.13	7	
	1000	40.21	1.72	11	

REFERENCES

1. Heard S. Lowry, III and Charles J. Fisher,"Line Parameter Measurements and Calculations of CO Broadened by Nitrogen at Elevated Temperatures," J. Quant. Spectrosc. Radiat. Transfer, vol.27, No. 6, pp. 585-591, 1982.

2. P. L. Varghese and R. K. Hanson, "Tunable Infrared Diode Laser Measurements of Line Strengths and Collision Widths of $^{12}C^{16}O$ at Room Temperature," J. Quant. Spectrosc. Radiat. Transfer, Vol. 24, pp. 479-489, 1980.

3. P. L. Varghese abd R. K. Hanson, "Collision Width Measurements of CO in Combustion Gases Using a Tunable Diode Laser," J. Quant. Spectrosc. Radiat. Transfer, Vol. 26, No. 4, pp. 339-347, 1981.

4. C.J. Tsao and B. Curnutte,"Line Widths of Pressure-Broadened Spectral Lines," J. Quant. Spectrosc. Radiat. Transfer, Vol. 2, pp. 41-91, 1962.

5. S. S. Penner, Quantitative Molecular Spectroscopy and Gas Emissivities, (Addison Wesley Publishing Company, Inc., Reading, Mass., 1959).

6. G. Birnbaum,"Microwave Pressure Broadening and Its Application to Intermolecular Forces," Advances in Chemical Physics, Vol. 12, 1967, pp. 487-548.

7. P. Varanasi and G. D. T. Tejwani, J. Quant. Spectrosc. Radiat. Transfer, Vol. 11, p. 255, 1971.

8. G. D. T. Tejwani, J. Quant. Spectrosc. Radiat. Transfer, Vol 12, p. 123, 1972.

9. J. A. Sell, J. Quant. Spectrosc. Radiat. Transfer, Vol. 23, p. 595, 1980.

10. P. L. Varghese and R. K. Hanson, J. Quant. Spectrosc. Radiat. Transfer, Vol. 24, p. 479, 1980.

11. D. White, D. E. Mann. P. N. Walsh, and A. Sommer, "Infrared Emission Spectrum of Gaseous HBO_2," Jour. Chem. Phys., Vol. 32, No. 2, pp. 488-492, 1960.

12. P. L. Hanst and V. H. Early, "Infrared Spectrum and Molecular Structure of B_2O_3," Jour. Chem. Phys., Vol. 42, No. 3, pp. 1097-1104, 1965.

13. D. J. Meschi, W. A. Chupka, and J. Berkowitz,"Heterogeneous Reactions Studied by Mass Spectrometry. I. Reaction of $B_2O_3(s)$ with $H_2O(g)$," Jour. Chem. Phys., Vol 33, No. 2, pp. 530-533, 1960.

14. W. J. Millerand R. K. Gould, "Electron Attachment Kinetics in Flames: Dissociative Attachment to HBO_2," Chem Phys. Letters, Vol. 38, No. 2, pp. 237-241, 1976.

15. Donald W. Boyer, "Shock-Tube Measurements of the Band Strengths of HBO_2 and OBF in the Short Wavelenths Infrared," J. Quant. Spectrosc. Radiat. Transfer, Vol. 24, pp.269-282, 1980.

16. Kirk-Othmer, Encyclopeida of Chemical Technology, third edition, Vol 4 (Interscience Publishers, John Wiley & Sons, New York, 1970).

17. Walter Gordy and Robert L. Cook, Microwave Molecular Spectra, (Interscience Publishers, John Wiley & Sons, New York, 1970).

18. Gerhard Herzberg, Infrared and Raman Spectra of Polyatomic Molecules, (D. Van Norstrad Company, Inc., Princeton, N. J., 1945).

1982 USAF-SCEEE SUMMER FACULTY RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION

FINAL REPORT

3-DIMENSIONAL GRID GENERATION WITH APPLICATIONS TO HIGH-PERFORMANCE AIRCRAFT

Prepared by:	David C. Wilson
Academic Rank:	Associate Professor
Department and University:	Department of Mathematics University of Florida
Research Location:	Flight Dynamics Laboratory, Wright-Patterson Air Force Base, Aerodynamics & Airframe Branch, Computational Aerodynamics Group
UASF Research Colleague:	Joe S. Shang
Date:	August 5, 1982
Contract No:	F49620-82-C-0035

3-DIMENSIONAL GRID GENERATION WITH APPLICATIONS TO HIGH PERFORMANCE AIRCRAFT

by

David C. Wilson

ABSTRACT

A general method of constructing 3-dimensional grids is developed using straight forward techniques from Trigonometry and Analytic Geometry. The method has the virture that explicit, yet simple, formulas are used. This procedure allows tight control of grid point location and mesh size. These ideas are then applied to the special case of a high performance aircraft with close-coupled canard.

ACKNOWLEDGEMENT

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education for providing him with the opportunity to spend a very worthwhile and interesting summer at the Flight Dynamics Laboratory, Wright-Patterson AFB, Dayton, Ohio.

Finally, he would like to thank Dr. Wilber Hankey and Dr. Joe Shang for their enthusiasm and guidance in directing this research. He would also like to thank Jeff Graham for his technical assistance.

I. INTRODUCTION

An important method for solving a system of partial differential equations on a given region subject to prescribed boundary conditions is to first transform the region to the computational domain (usually the unit square or unit cube), numerically solve this new system in the rectilinear coordinates, and then transform back to the physical domain to get the final solution. Since the flow of a fluid is described by the Navier-Stokes equations, special cases of this problem arise continually in fluid dynamics. In this paper the author will not be concerned with differential equations, but instead will focus his attention on the problem of constructing suitable grid transformation.

Using various conformal mapping techniques, as well as the algebraic[1,2], elliptic[3], and hyperbolic[4] methods, two dimensional grids can usually be constructed in a straight forward and systematic way. The problem in three dimensions is not so simple. Not only is it much more difficult to describe an arbitrary region (the boundaries are no longer simple closed curves), but the methods themselves become more difficult to apply.

The grids generated by the elliptic and hyperbolic methods are given implicitly as solutions of second order partial differential equations. These two methods have the virture that the transformations are guaranteed to have smooth second derivatives. Given reasonable conditions on the region it is known from the basic theory of harmonic functions that there will always exist a unique transformation which is elliptic. Moreover, this transformation will have the virtue that it is a diffeomorphism if it is one-to-one on the boundary.

The algebraic method is an explicit method because formulas are found which smoothly interpolate the region between two surfaces. Since the formulas are given explicitly there is very tight control of the spacing of the grid lines and the size of mesh. However, care must be exercised to ensure that the result is one-to-one smooth, and nearly orthogonal.

The basic method of this paper is to visualize a transformation which "pushes" (or "molds") an aircraft into the boundary of the unit cube in the smoothest and least provocative way possible. Next, as in the algebraic method, explicit formulas are derived which will mathematically describe this push. The formulas will be simple in the sense that only analytic geometry and trigonometry are required for their formulation. Since the transformation is given explicitly, very tight control can be maintained over the spacing of the grid points. Also, derivatives are easy to compute. However, as in the algebraic method, we must exercise care to be sure that the transformation is one-to-one. Since the transformation is patched together from a number of partial mappings, care must even be exercised to ensure that the transformation is continuous.

To illustrate this method a 3-dimensional grid is constructed for the aircraft sketched in Figure 1. A model of this aircraft was tested in three different configurations by W. Calarese[5] at the Air Force Wright Aeronautical Laboratories at Wright-Patterson Air Force Base, Ohio. The first configuration was with the canard missing, the second with the canard in the middle position, and the third with the canard in the raised position. Since the aircraft tested is symmetric with respect to both the up-down and left-right directions, the grids will be restricted to one quarter of the aircraft. Even though the case when the canard and airfoil are on the same level is geometrically no more difficult than the case with the canard missing, it is left undone because the author is concerned that

the method indicated here would lead to problems when solving the Navier-Stokes equations.

II. OBJECTIVES

The first goal of this research is to develop new methods for constructing 3-dimensional grid transformations which will be useful in the study of the flowfield around a high speed aircraft. We did not attempt to solve the general problem, but rather we developed a grid for one specific aircraft.

III. DEFINITIONS AND BASIC CONCEPTS

Since the method of 3-dimensional grid generation is to construct a diffeomorphism from the computational space to the physical space and then find the images of the constant ζ, η, and ξ surfaces, it is useful to make the following definitions.

Definition 1. If $H:X\to Y$ is a function between sets X and Y, then H is said to be one-to-one if $H(x) \neq H(x')$ whenever $x \neq x'$.

Definition 2. If $H:X\to Y$ is a function between sets, then H is said to be onto if for every $y \in Y$ there is an $x \in X$ such that $H(x) = y$. If $H:X\to Y$ is both one-to-one and onto, then the inverse function $H^{-1}:Y\to X$ is a well-defined function which is also one-to-one and onto.

Let $E^3 = \{(x, y, z) | x, y, z \text{ are real numbers}\}$ denote Euclidean 3-space, and let $I^3 = \{(x, y, z) \in E^3 | 0 \leq x, y, z \leq 1\}$.

Definition 3. If X and Y are subsets of E^3 and $H:X\to Y$ is both one-to-one and onto, then H is called a homeomorphism if both H and H^{-1} are continuous.

Definition 4. If X and Y are subsets of E^3, then a function $F:X\to Y$ is called smooth if the first and second partial derivatives of the coordinate functions exist and are continuous. A smooth homeomorphism H with smooth inverse H^{-1} is called a diffeomorphism.

The reason for introducing this terminology is that a 2-dimensional grid transformation is nothing more than a diffeomorphism from the unit square into the plane. A 3-dimensional grid is a diffeomorphism defined from I^3 into E^3.

IV. A GRID FOR THE MODEL MISSING THE CANARD

To construct a grid for the model when the canard is missing we will first find a homeomorphism $R(\xi, \eta) = (X(\xi, \eta), Y(\xi, \eta))$ from the (ξ, η)-plane into the (x, y)-plane which slides a rectangle under the airfoil. Figure 2 provides a diagram of $R(\xi, \eta)$. If the function $Z(\xi, \eta)$ is the height of the airfoil above the point $R(\xi, \eta)$ then the homeomorphism

$$S(\xi, \eta) = (X(\xi, \eta), Y(\xi, \eta), Z(\xi, \eta)) \tag{1}$$

will stretch the rectangle over the top of the airfoil. The final grid transformation $T(\xi, \eta, \zeta)$ will wrap the $\eta = 0$ plane around the surface of the fuselage. The transformation $T(\xi, \eta, \zeta)$ is sketched in Figure 3. For computational purposes a long narrow shaft has been added to the nose.

To describe the transformations $R(\xi, \eta)$ and $Z(\xi, \eta)$ we need to write the first quadrant of the (ξ, η)-plane, denoted by D, as the union of six subsets. If $0 < a < b < c < d < e$ and $w > 0$, then let $D_1 = \{(\xi, \eta) \in D | \xi \leq a\}$, $D_2 = \{(\xi, \eta) \in D | a \leq \xi \leq b\}$, $D_3 = \{(\xi, \eta) \in D | b \leq \xi \leq c\}$, $D_4 = \{(\xi, \eta) \in D |$ $c \leq \xi \leq d$ and $0 \leq \zeta \leq w\}$, $D_5 = \{(\xi, \eta) \in D | c \leq \xi$ and $w \leq \eta\}$, and $D_6 = \{(\xi, \eta)$ $\in D | d \leq \xi$ and $0 \leq \eta \leq w\}$. Note that $D = D_1 \cup D_2 \cup \ldots \cup D_6$.

Let $0 < \bar{a} < \bar{b}_1 < \bar{b}_2 < \bar{c} < \bar{d} < \bar{e}$, $f < g$, and $r_0 < r_1 < r_2$. The quantity r_0 denotes the radius of the shaft, r_1 denotes one half the width of the fuselage, and r_2 denotes one half the height of the fuselage. (For this model $r_1 = 3.21$ cm and $r_2 = 5.08$ cm). If $x_0 \geq 0$, then define $r(x_0)$ to be the radius of the circle generated when the plane $x = x_0$ is intersected with the fuselage. For $\bar{a} \leq x \leq \bar{b}$, $r(x)$ is a parabola so that constants A, B, and C

can be found such that $r(x) = Ax^2 + Bx + C$ and $r(\bar{a}) = r_0$, $r(\bar{b}_1) = r_1$ and $r(\bar{b}_2) = r_2$.

The restriction of the function R to the set D_i will be denoted by R_i. (i.e. $R_i = R|D_i$). If the distance from the fuselage to the wingtip is denoted by $\bar{w}$ and if $(\xi, \eta) \in D_1$, then define

$$R_1(\xi, \eta) = (\frac{\bar{a}}{a}\xi, \frac{\bar{w}}{w}\eta + r_0). \tag{2}$$

If $(\xi, \eta) \in D_2$, define

$$R_2(\xi,\eta) = (\bar{a} + \frac{\bar{b}-\bar{a}}{b-a}(\xi - a)), \frac{\bar{w}}{w}\eta + r(\bar{a} + \frac{\bar{b}_1-\bar{a}}{b-a}(\xi - a)) \tag{3}$$

Let $P_1 = (\bar{b}_1, r_1 + \bar{w})$, $P_2 = (\bar{b}_1, r_1)$, $P_3 = (\bar{c}, r_1)$, and $P_4 = (f, r_1 + \bar{w})$. If $(\xi, \eta) \in D_3$, then define

$$R_3(\xi,\eta) = \eta/w\,(1 - \frac{\xi-b}{c-b})\,P_1 + (1 - \frac{\xi-b}{c-b})(1 - \eta/w)\,P_2 + (1 - \eta/w)(\frac{\xi-b}{c-b})\cdot$$

$$P_3 + \eta/w\,\frac{\xi-b}{c-b}\,P_4. \tag{4}$$

If $Q_1 = (\bar{d}, r_1)$, $Q_2 = (g, r_1 + \bar{w})$, and (ξ, η) D_4, then define

$$R_4(\xi,\eta) = \eta/w\,(1 - \frac{\xi-c}{d-c})\,P_4 + (1 - \eta/w)(1 - \frac{\xi-c}{d-c})\,P_3 + (1 - \eta/w)\,\frac{\xi-c}{d-c}\,Q_1$$

$$+ (1 - \eta/w)\,\frac{\xi-c}{d-c}\,Q_2. \tag{5}$$

Define $R_5(\xi, \eta)$ and $R_6(\xi, \eta)$ on the sets D_5 and D_6 respectively by the rules

$$R_5(\xi, \eta) = R_3(c, \eta) + R_4(\xi, w) - (f, r_1 + \bar{w}) \tag{6}$$

and

$$R_6(\xi, \eta) = R_4(d, \eta) + R_5(\xi, w) - (g, r_1 + \bar{w}). \tag{7}$$

Since R_i and R_j agree on the sets $D_i \cap D_j$ and since it can be easily checked that each R_i is a diffeomorphism, the transformation $R = R_1 \cup R_2 \cup \ldots \cup R_6$ is a homeomorphism. However, since the derivatives of R_i and R_j do not agree on the set $D_i \cap D_j$, R is <u>not</u> a diffeomorphism. Since in applications the derivatives will be taken numerically, the hope is that the discontinuities of the first and second derivatives should be smoothed out a bit.

We are now ready to define the mapping $Z(\xi, \eta)$. If $(\xi, \eta) \in D - D_4$, then define $Z(\xi, \eta) = 0$. To define $Z(\xi, \eta)$ on D_4 we must first describe functions $\rho(y)$, $\alpha(y)$, $\beta(y)$, $\gamma(y)$, and $\delta(y)$. The relationships between these quantities are diagramed in Figure 4. If $\alpha(r_1)$ is the height of the airfoil above the point $P = (\frac{\bar{c} + \bar{d}}{2}, r_1)$, then by the specifications of the model $\alpha(r_1)$ is 6% of one half of the maximum chord length or $\alpha(r_1) = (.03)\cdot(\bar{d} - \bar{c})$. If $\alpha(r_1 + \bar{w})$ denotes the height of the airfoil at the wingtip, then $\alpha(r_1 + \bar{w}) = (.02)(g - f)$. However, to keep the formulas and computations to a minimum we will assume $\alpha(r_1 + \bar{w}) = 0$. The quantities $\rho(y)$, $\alpha(y)$, $\beta(y)$, $\gamma(y)$, and $\delta(y)$ are defined by the following rules.

1. The radius of the circular arc obtained when the plane $y = y_0$ is intersected with the airfoil is denoted by $\rho(y_0)$.
2. The maximum height of the airfoil in the plane $y = y_0$ is denoted by $\alpha(y_0)$.
3. The quantity $\beta(y)$ is defined by $\beta(y) = \rho(y) - \alpha(y)$.
4. The quantity $\gamma(y_0)$ denotes one half of the diameter of the set obtained when the plane $y = y_0$ is intersected with the airfoil or

$$\gamma(y_0) = (1 - \frac{y_0 - r_1}{\bar{w}}) \frac{\bar{d} - \bar{c}}{2} + \frac{g - f}{2} (\frac{y - r_1}{\bar{w}}) \qquad (8)$$

5. The formula for $\delta(y_0)$ is

$$\delta(y_0) = \left(1 - \frac{y_0 - r_1}{\bar{w}}\right) \frac{\bar{c}+\bar{d}}{2} + \frac{y_0 - r_1}{\bar{w}} \frac{f+g}{2} \tag{9}$$

(i.e. $\delta(y_0)$ is the x-axis midpoint of the intersection of the plane $y = y_0$ and the middle of the airfoil.)

Since the highest point of the airfoil in the plane $y = y_0$ lies on the line through the points $(\frac{c+d}{2}, r_1, \alpha(r_1))$ and $(\frac{f+g}{2}, r_1 + \bar{w}, 0)$, $\alpha(y)$ is given by the formula $\alpha(y) = \alpha(r) - \alpha(r)\frac{y-r_1}{\bar{w}}$. Since $\rho^2(y) = \beta^2(y) + \gamma^2(y)$ and $\rho(y) = \alpha(y) + \beta(y)$, $\rho(y) = \frac{\alpha^2(y) + \gamma^2(y)}{2 \cdot \alpha(y)}$. If $\bar{c} \leq x \leq \bar{d}$ and (x, y, z) denotes the point on the airfoil above the point $(x, y, 0)$, then

$$z = -\beta(y) + \sqrt{\rho^2(y) - (x - \delta(y))^2} \tag{10}$$

If $(\xi, \eta) \in D_4$, then define

$$Z(\xi, \eta) = -\beta(Y(\xi, \eta)) + \sqrt{\rho^2(Y(\xi, \eta)) - (X(\xi, \eta) - \delta(Y(\xi, \eta)))^2} \tag{11}$$

To describe the fuselage we can think of it being manufactured in two stages. First a paraboloid of revolution (the nose) is glued to a round circular (the body). Second, this solid is placed in a saw mill with two vertical blades which shave an equal amount from each side. If $h(x_0)$ denotes the height of the shaved side of the fuselage in the $x = x_0$ plane, then the vertical side is the region $\{(x, r_1, z) | 0 \leq z \leq h(x)$ and $x \geq \bar{b}_1\}$. If $x \leq \bar{b}_1$, then $h(x) = 0$. If $x \geq \bar{b}_1$, then the Pythagoren Theorem can be applied to show that $h(x) = \sqrt{r^2(x) - r_1^2}$. Figure 5 gives a cross-sectional view of the fuselage. If $x \geq \bar{b}_2$, then $r(x) = r_2$.

We are now in a position to define the grid transformation $T(\xi, \eta, \zeta)$. Let $v > 0$, $h = h(\xi, \eta) = h(X(\xi, \eta))$, $r(\xi, \eta) = r(X(\xi, \eta))$ and $v(\xi, \eta) = \frac{h(\xi, \eta)}{r(\xi, \eta)} v$. If $0 \leq \zeta \leq v(\xi, \eta)$, then define

$$T(\xi,\eta,\zeta) = \left(X(\xi,\eta), Y(\xi,\eta), Z(\xi,\eta) + \frac{\zeta}{v(\xi,\eta)} (h(\xi,\eta) - Z(\xi,\eta))\right) \tag{12}$$

Since $T(\xi,\eta, 0) = (X(\xi,\eta), Y(\xi,\eta), Z(\xi,\eta))$, the $\zeta = 0$ plane is mapped into

the $z = 0$ plane and onto the top of the airfoil. Note also that the region $\{(\xi, 0, \zeta) | 0 \leq \zeta \leq v(\xi, 0)\}$ is mapped onto the vertical (or shaved) side of the fuselage. Note further the $T(\xi, \eta, v(\xi, \eta)) = (X(\xi, \eta), Y(\xi, \eta), h(\xi, \eta))$.

If $v(\xi, \eta) \leq \zeta \leq v$, then define

$$\theta = \theta(\xi,\eta,\zeta) = \left(1 - \frac{\zeta - v(\xi,\eta)}{v - v(\xi,\eta)}\right) \tan^{-1}\left(\frac{h(\xi,\eta)}{Y(\xi,\eta)}\right) + \pi/2 \frac{\zeta - v(\xi,\eta)}{v - v(\xi,\eta)} \text{ and} \tag{13}$$

$$T(\xi,\eta,\zeta) = \left(X(\xi,\eta), \sqrt{Y^2(\xi,\eta) + h^2(\xi,\eta)} \cdot \cos\theta, \sqrt{Y^2(\zeta,\eta) + h^2(\xi,\eta)} \cdot \sin\theta\right). \tag{14}$$

V. THE RAISED CANARD

The goal of this section is to describe a grid transformation for the model with the canard in the raised position. Since the design of the canard is similar to that of the airfoil, we again must find functions Rc, Sc, ρc, αc, etc. which have the same relationships as the corresponding functions defined for the airfoil. (Any quantity subscripted with the letter c will be associated with the canard.)

Let $h_c = h_c(x) = h_c(X(\xi, \eta))$ be any (preferably smooth) function which represents the height of the middle of the canard when $\bar{c}_c \leq x \leq \bar{d}_c$ and which has the property that $h(x) < h_c(x) < r(x)$ for all x. Let r_3 be the distance from the $y = 0$ plane to the canard-fuselage intersection. Find the mapping $S(\xi,\eta,\zeta) = (X(\xi,\eta), Y(\xi,\eta), Z(\xi,\eta)$ which locates the airfoil in 3-space. Next apply the method of Section IV to find a mapping $Rc(\xi,\eta) = (X_c(\xi,\eta), Y_c(\xi,\eta))$ which runs the ξ-axis around the edge of the fuselage and maps the rectangle $D_{c,4}$ onto the "translated shadow" of the canard in the (x, y) plane. The "translated shadow" is the "shadow" of the actual canard translated $r_1 - r_3$ units in the positive y direction.

If $0 \leq \zeta \leq v(\xi,\eta)$, then define $T(\xi,\eta,\zeta)$ by the formula

$$T(\xi,\eta,\zeta) = \left(1 - \frac{\zeta}{v(\xi,\eta)}\right) S(\xi,\eta) + \frac{\zeta}{v(\xi,\eta)} (Xc(\xi,\eta), Yc(\xi,\eta), h(\xi,\eta)) \tag{15}$$

(If $v(\xi,\eta) = 0$, then let $T(\xi,\eta,\zeta) = S(\xi,\eta)$.)

As in the case of the airfoil a cross section of the canard is described by two circular arcs whose maximum heights are over the middle. The maximum height in the plane $y = r_3$ is 6% of the chord length at the fuselage or $\alpha_c(r_3) = (0.3)\ (\bar{d}_c - \bar{c}_c)$. Even though the maximum height at the tip of the canard is 4% of the chord length at the tip, we will again assume that the tip is sharp. (i. e. $\alpha_c(r_3 + \bar{w}_c) = 0$.)

As in the case of the airfoil the formula which gives the height of the upper and lower surface of the canard above the (x, y)-plane is given by

$$z = h_c \pm \alpha_c(y) - \rho_c(y) + \sqrt{\rho_c^2\ (y) - (x - \delta_c(y))^2} \tag{16}$$

We now want to derive formulas for the lower and upper surfaces of the canard in terms of ξ and η. Let L(t) be the line through the points

$$P_1 = (X_c(\xi,\eta),\ \bar{w}_c + \sqrt{r_2^2 - X_c^2\ (\xi,\eta) - h_c^2},\ h_c) \tag{17}$$

$$P_2 = (X_c(\xi,\eta),\ \sqrt{r_2^2 - X_c^2(\xi,\eta) - h_c^2},\ h_c - \alpha_c(r_3))$$

If $L(t) = P_1 + (P_2 - P_1)\cdot t$, then let t_1 be the smallest point such that $L(t_1)$ is on the cylinder $y^2 + z^2 = r_2^2$. Let $M(t) = P_1 + (P_2' - P_1)\cdot t$ be the line through the points P_1 and

$$P_2^1 = (X_c(\xi,\eta),\ \sqrt{r_2^2 - X_c^2(\xi,\eta) - h_c^2},\ h_c + \alpha_c(r_3)). \tag{18}$$

Let t_2 be the smallest point such that $M(t_2)$ is on the cylinder $y^2 + z^2 = r_2^2$. If we let $Z_c^{(1)}(\xi,0)$ denote the z-coordinates of $L(t_1)$, then define

$$Z_c^1(\xi,\eta) = (1 - \eta/w)\ Z_c^{(1)}\ (\xi,0) + \eta/w\ h_c\ . \tag{19}$$

Using the line M(t) define $Z_c^{(2)}(\xi,\eta)$ by the formula given by equation (19).

If $(\xi,\eta)\ D_{c,4}$ and $\zeta \geq \frac{h}{r_2}\ v$, then define

$$\theta_1 = \tan^{-1}\left(\frac{h(\xi,\eta)}{Y_c(\xi,\eta)}\right) \tag{20}$$

Let t_0 be the smallest point such that the distance between $L(t_0)$ and $(X_c(\xi,\eta),\ 0,\ 0)$ is $\sqrt{X_c^2(\xi,\eta) + Y_c^2(\xi,\eta) + h^2}$. If $Y_L(t_0)$ and $Z_L(t_0)$ denote the y and z coordinates of $L(t_0)$, respectively, then define

$$\theta_2 = \tan^{-1}\left(\frac{Z_L(t_0)}{Y_L(t_0)}\right) \tag{21}$$

If $\frac{h}{r_2}\ v < \zeta < v$ and Θ is given by the formula

$$\Theta = \theta_1 + \frac{\theta_2 - \theta_1}{v - \frac{h}{r_2}\ v}\left(\zeta - \frac{h}{r_2}\ v\right) \tag{22}$$

then define

$$T(\xi,\eta,\zeta) = \left(X_c(\xi,\eta),\ \sqrt{X_c^2(\xi,\eta) + Y_c^2(\xi,\eta) + h^2}\cdot\cos\Theta,\right.$$
$$\left.\sqrt{X_c^2(\xi,\eta) + Y_c^2(\xi,\eta) + h^2}\cdot\sin\Theta\right). \tag{23}$$

Let t_0' be the smallest point such that the distance between $M(t_0')$ and $(X_c(\xi,\eta),\ 0,\ 0)$ is $\sqrt{X_c^2(\xi,\eta) + Y_c^2(\xi,\eta) + h^2}$. If $Y_M(t_0')$ and $Z_M(t_0')$ denote the y and z coordinates of $M(t_0')$, respectively, then define $\theta_3 = \tan^{-1}\left(\frac{Z_M(t_0')}{Y_M(t_0')}\right)$. If $v < \zeta < 2v$ and $\Theta = \theta_3 + \frac{\pi/2 - \theta_3}{v}(\zeta - v)$, then define $T(\xi,\eta,\zeta)$ as in equation (23).

We have now completed the definition of the transformation $T(\xi,\eta,\zeta)$. Even though it can be easily checked that T is a homeomorphism off the set $A = \{(\xi,\eta,\zeta)\,|\,(\xi,\eta) \in D_{c,4}$ and $\zeta = h\}$ it will fail to be smooth where it is patched together. (Note that the transformation T is not even well-defined on the set A so that care must be taken here!)

VI. RECOMMENDATIONS

Even though the grid transformations we have constructed are very well behaved homeomorphisms, it would be better if they were smooth everywhere. In general we would like our methods to combine the virtues of both the algebraic and the PDE methods. Namely, we would like our transformations to be easily constructed, smooth, and to provide excellent grid control. The next step is to improve our grids by making them smooth where they are patched together. I propose to do this by developing and applying some spline function techniques for two and three dimensions. These methods would then be integrated into the procedure described in this paper to give an algorithm for solving a large class of problems of this type. While engineers are not much interested in abstract mathematical statements, the art of grid generation would be greatly enhanced if an algorithm could be found which would prove the following theorem from 3-space topology: If $h: \partial I^3 \to E^3$ is a piecewise smooth homeomorphism from the boundary of I^3 into E^3, then there is a homeomorphism $H: I^3 \to E^3$ which is smooth inside I^3 and agrees with h on ∂I^3. If an efficient algorithm could be found to find the extension H, the area of grid generation would be advanced because both the manpower needed to develop the grid and the computational time needed in applying the grid to a specific problem could be reduced substantially.

The long term goal is to develop a classification scheme which would divide all aircraft configurations into one of a few different types. The idea would be to then have an algorithm which would automatically develop a grid for each type. At this point we could consider the problem of 3-D grid generation to be manageable.

REFERENCES

1. P.R. Eiseman and R. E. Smith, "Mesh Generation Using Algebraic Techniques," Numerical Grid Generation Techniques, NASA Conference Publication 2166, pp. 73-120, 1980.
2. R. E. Smith, "Algebraic Grid Generation," (Preprint).
3. J. F. Thompson, F. C. Thames, and C. W. Mastin, "Automatic Numerical Generation of Body-Fitted Curvilinear Coordinate System for Field Containing Any Number of Arbitrary Two-Dimensional Bodies," Journal of Computational Physics, Vol. 15, pp. 299-319, 1974.
4. J. L. Steger and D. S. Chaussee, "Generation of Body Fitted Coordinates Using Hyperbolic Partial Differential Equations," (Preprint).
5. W. Calerese, "Close-Coupled Canard-Wing Vortex Interaction and Reynolds Stress Acquisition," (Preprint).

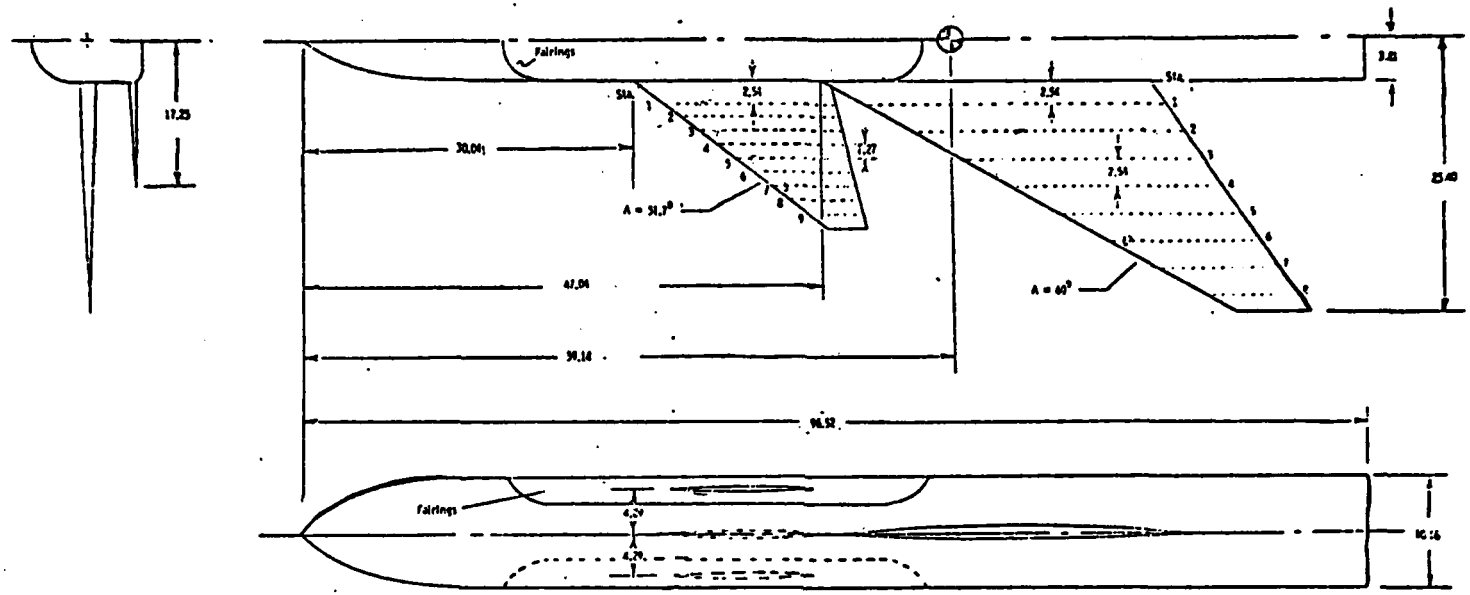


Figure 1.- Sketch of Model. All dimensions in centimeters.

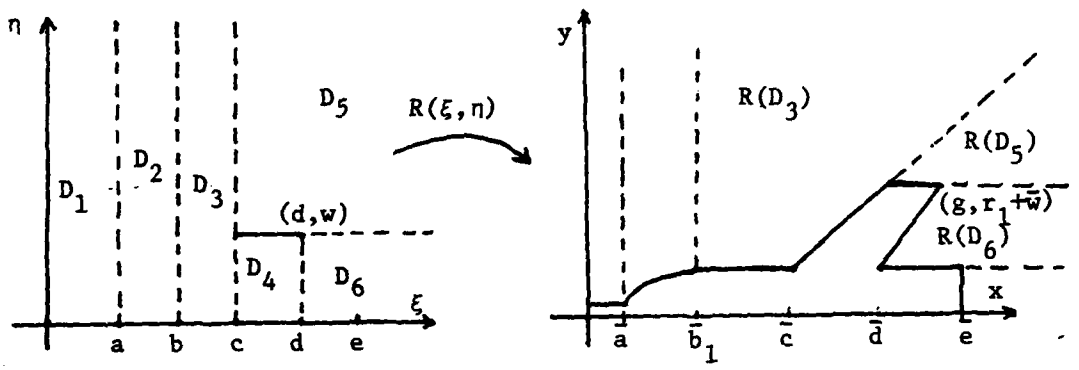


Figure 2.- The Transformation $R(\xi,\eta)$.

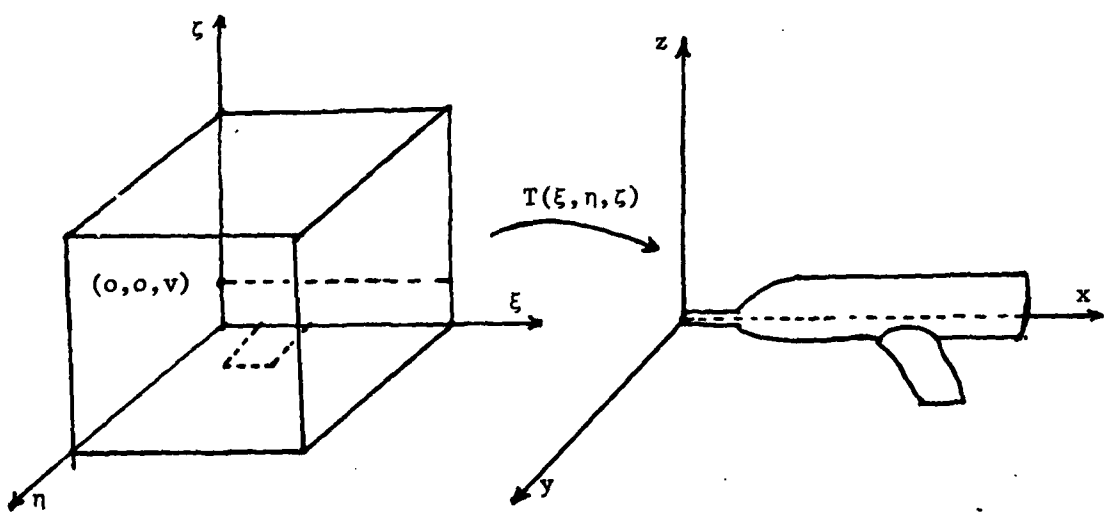


Figure 3.- The Transformation $T(\xi,\eta,\zeta)$.

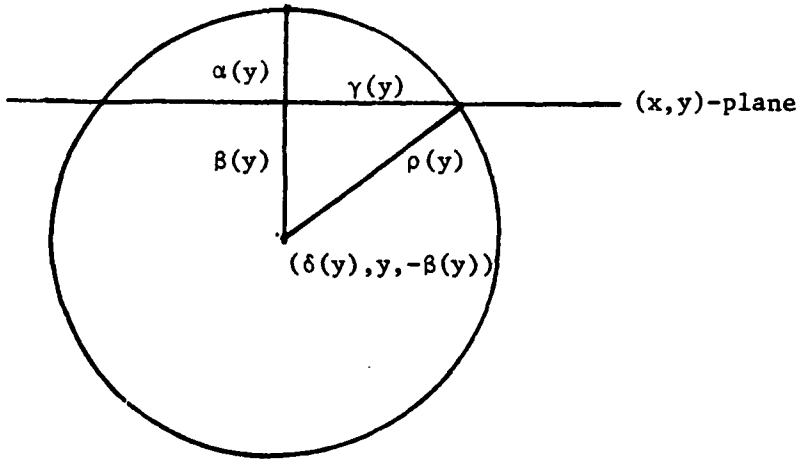


Figure 4.- A View of the Airfoil.

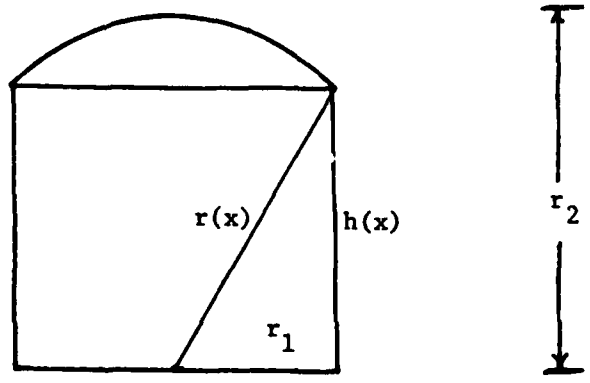


Figure 5.- A Cross Section of the Fuselage.

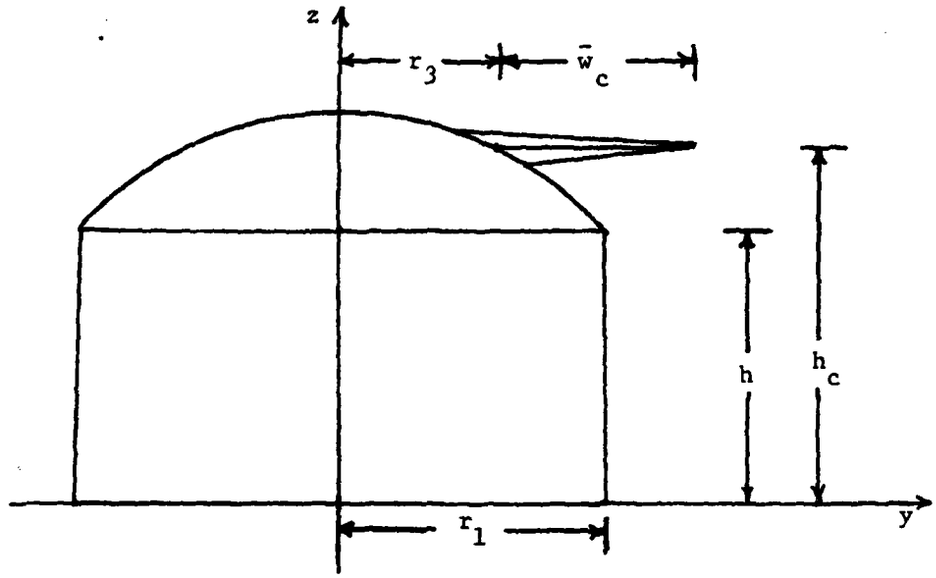


Figure 6.- The Canard-fuselage Intersection.

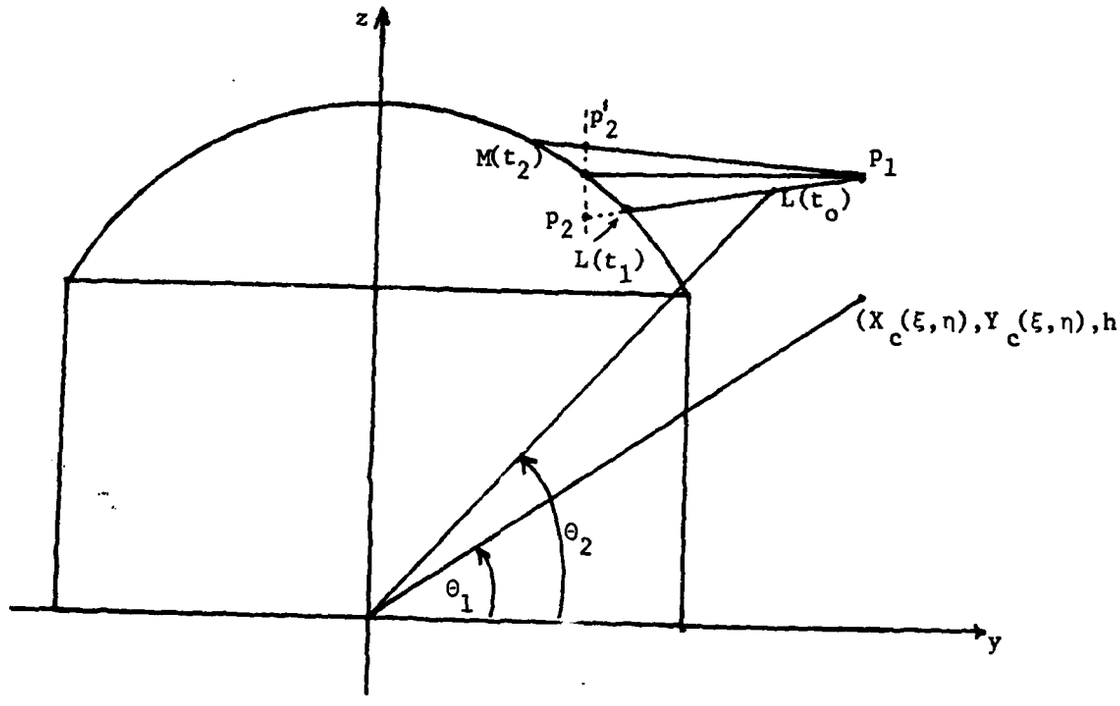


Figure 7.- The Key Points on the Canard.

1982 USAF-SCEEE Summer Faculty Research Program

Sponsored by the

Air Force Office of Scientific Research

Conducted by the

Southeastern Center for Electrical Engineering Education

Final Report

Time Domain Analysis and Synthesis of Robust Controllers for Large Scale LQG Regulators

Prepared by:	Dr. Rama Krishna Yedavalli
Academic Rank:	Assistant Professor
Dept. and Univ.:	Department of Mechanical Engineering Stevens Institute of Technology
Research Location:	Flight Dynamics Laboratory, Flight Control Division, Control Analysis Group
USAF Research Colleague:	Dr. Siva S. Banda
Date:	September 7, 1982
Contract No.:	F49620-82-C-0035

Acknowledgement

The author would like to thank the Air Force Systems Command, the Air Force Office of Scientific Education for providing him with the opportunity to spend a very fruitful and rewarding summer at the Flight Dynamics Laboratory, Wright-Patterson AFB, Ohio. He would like to acknowledge the laboratory, in particular the Control Analysis Group, for its hospitality and excellent working conditions.

The author would like to thank Dr. Siva S. Banda for his interest, guidance and assistante in this research. Thanks are also due to Dr. Ronald Anderson, Mr. Frank George, Mr. Dave Bowser for their encouragement. Finally he wishes to acknowledge many helpful and stimulating discussions with Dr. Robert C. Schwanz, Major Robert Rennard, Lt. D.B. Ridgeley.

Time Domain Analysis and Synthesis of Robust Controllers for Large Scale LQG Regulators

by

Rama Krishna Yedavalli

Abstract

The aspect of 'Robustness' in the analysis and synthesis of time domain large scale Linear Quadratic Gaussian (LQG) regulators is addressed. Both 'Stability Robustness' and 'Performance Robustness' are combinedly considered to meet stability and performance requirements. The perturbations which cause instability in the system are viewed as consisting of two types of modeling errors: namely parameter variations and model/controller truncation. Conditions are derived for 'Stability Robustness' in terms of the singular values of the closed loop system matrix and the perturbation matrix for various cases of perturbations. Simple measures of 'Stability Robustness' and 'Performance Robustness' are derived and based on these measures, a control design algorithm is presented that achieves a satisfactory trade-off between stability and performance. Applications in the field of Large Space Structure (LSS) Control are discussed. Suggestions for further research are offered.

I. Introduction:

It is well known that the inaccuracies in the mathematical models of physical systems can severely compromise the resulting control designs. The errors associated with mathematical models of physical systems may be broadly categorized as i) parameter errors, ii) truncated models (errors in model order), iii) neglected or incorrectly modeled external disturbances and iv) neglected nonlinearities. It is the inevitable presence of these errors in the model used for design that eventually limits the performance attainable from the control system designs produced by either classical (frequency domain) or modern (time domain) control theory. The problem of model errors is more critical, in general, for large scale Linear Quadratic Gaussian (LQG) regulator problems and in particular, for Large Space Structure (LSS) control, an application of extreme importance to the U.S. Air Force. The fundamental problem of LSS control is the control of a large dimensional system with a controller of much smaller dimension (model/controller truncation) compounded with modal data uncertainty (parameter errors). In the light of these observations, it is evident that 'robustness' is an extremely desirable (sometimes, necessary) feature of any feedback control design proposed for LSS control. 'Robustness' studies of Large Scale LQG regulators is the central theme, of the present report.

For our present purposes a 'robust' control design is that design which behaves in an 'acceptable' fashion (i.e. satisfactorily meets the system specifications) even in the presence of modeling errors. Since the system specifications could be either in terms of stability and/or performance (regulation, time response, etc.) we can conceive two types of robustness, namely 'Stability Robustness' and 'Performance Robustness'. Limiting our

attention in this research to 'parameter errors' and 'model/controller truncation' as the two types of modeling errors that may cause instability (in performance degradation) in the system, we formally define 'stability robustness' and 'performance robustness' as follows:

'Stability Robustness': Maintaining closed loop system stability in the presence of modeling errors mainly parameter variations and model/controller truncation.

'Performance Robustness': Maintaining satisfactory level of performance in the presence of modeling errors mainly parameter variations and model/controller truncation.

Implicit in the definition of 'Performance Robustness' is the requirement of 'stability' for performance robustness studies. However performance robustness studies which require (or assume) stability may only have limited application. On the other hand concentrating design efforts on 'stability' alone is not prudent because the performance requirements may not be met by that design. Thus simultaneous consideration of stability and performance in the design process is more appropriate. Most of the current published literature addresses either the 'stability robustness' aspect or the 'performance robustness' aspect separately. Most of the interesting work on 'stability robustness' is done in frequency domain using singular value decomposition[1-9] while much of the useful research on 'performance robustness' is carried out in time domain using sensitivity approaches.[10-12] Design studies that combine both stability robustness and performance robustness have been scarce. In this research the objective is to consider both stability robustness and performance robustness simultaneously and this is elaborated in the next section.

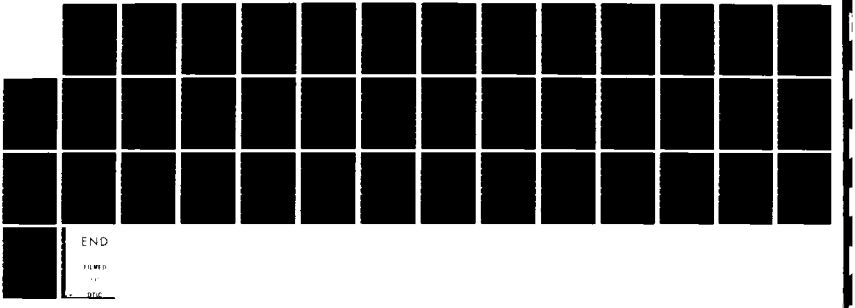
II. Objectives:

It is evident from the above discussion that a realistic design needs to consider both 'stability' and 'performance' simultaneously in the design process, especially in the light of stringent requirements on performance like in flexible spacecraft (aircraft) pointing control problems. Thus the main objective of this research was to develop a design algorithm that incorporates both 'stability robustness' and 'performance robustness' aspects into the problem formulation. However, in doing so, one must realize the fact many sensitivity studies brought out that one may not be able to achieve both stability and performance to desirable extents simultaneously by any given design. A trade-off between stability and performance is to be expected. It is this perspective that the proposed design algorithm takes in incorporating both stability robustness and performance robustness into the design formulation and this is done in time domain LQG regulator format. The specific objectives of the project can now be summarized as follows:

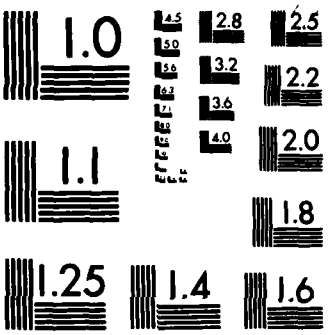
1. To characterize the perturbation matrices in terms of the two modeling errors, namely 'parameter errors', and 'model/controller truncation' and to develop condition for stability in terms of the singular values of the perturbation matrix and a nominally stable system matrix (this concept, which is carried out here in time domain, runs somewhat parallel to the frequency domain approach of [1]).

2. To develop simple measures of stability robustness and performance robustness and use them to present a design algorithm for robust controller design which achieves a trade-off between them, employing the standard time

AD-A130 770 USAF/SCEEE SUMMER FACULTY RESEARCH PROGRAM (1982) 11/11
RESEARCH REPORTS VOLUME..(U) SOUTHEASTERN CENTER FOR
ELECTRICAL ENGINEERING EDUCATION INC S..
UNCLASSIFIED W D PEELE ET AL. OCT 82 AFOSR-TR-83-0614 F/G 5/1 NL



END



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

domain LQG regulator format.

3. To illustrate the design procedure by simple examples, suggesting the application to an LSS problem as future work. With this perspective, the report is organized as follows: Section III develops the conditions for stability of the closed loop system for the two types of modeling errors separately and then for the combined case. A simple measure of stability robustness is then presented. Section IV treats the case of performance robustness. Section V presents a design algorithm for robust controllers which is illustrated by simple examples. Finally Section VI offers some concluding remarks and suggestions for follow-on research.

III. Time Domain 'Stability Robustness' Analysis of LQG Regulators with Reference to LSS Models:

3.1 System Description:

In sequel, in the entire text of the report, the following linear continuous system is considered for analysis/design purposes. Attention is restricted to the time-invariant case. The differential and algebraic equations that govern the system are given by

$$\dot{x}(t,\bar{p}) = A(\bar{p})\, x(t,\bar{p}) + B(\bar{p})\, u(t,\bar{p}) + D(\bar{p})\, w(t) \tag{3.1a}$$

$$y(t,\bar{p}) = C(\bar{p})\, x(t,\bar{p}) \tag{3.1b}$$

$$z(t,\bar{p}) = M(\bar{p})\, x(t,\bar{p}) + v(t) \tag{3.1c}$$

where the state vector x is nx1, the control u is mx1, the external disturbance w is qx1, the output y (the k variables we wish to control) is kx1 and the measurement vector z is ℓx1. Accordingly the matrix A is of dimension nxn, B is nxm, D is nxq, C is kxn and M is ℓxn. The initial condition x(0)

is assumed to be a zero mean, gaussian random vector with variance Σ_o, i.e.

$$E[x(0)] = 0, \quad E[x(0)\, x^T(0)] = \Sigma_o \tag{3.2}$$

Similarly the process noise w(t) and the measurement noise v(t) are assumed to be zero mean white noise processes with gaussian distribution having constant covariances W and V respectively, i.e.

$$E[w(t)] = E[v(t)] = 0 \tag{3.3}$$

$$E\begin{bmatrix} w(t) \\ v(t) \end{bmatrix} \begin{bmatrix} w^T(\tau) & v^T(\tau) \end{bmatrix} = \begin{bmatrix} W & 0 \\ 0 & \rho_e V_o \end{bmatrix} \delta(t-\tau), \quad \text{scaler } \rho_e > 0, \quad V = \rho_e V_o$$

The system matrices A, B, C, D and M and consequently x, u, y and z are assumed to be continuous functions of constant but uncertain parameters $p_1, p_2, \ldots . p_r$ with nominal value $p_i = \bar{p}_i (i = 1,2,\ldots . r)$. Let the vector of uncertain parameters be denoted by p where

$$p^T = (p_1, p_2, \ldots . p_r) \tag{3.4}$$

The variation of the parameter vector p,

$$\delta p = p - \bar{p} \tag{3.5}$$

is assumed to be a zero mean random vector with variance

$$E[(\delta p)(\delta p)^T] = \text{Diag}[\sigma^2_{p_1}, \ \sigma^2_{p_2} \ldots . \sigma^2_{p_r}] \tag{3.6}$$

i.e. $\sigma^2_{p_i}$ represents the variance of the uncertain parameter p_i.

Let the above system be evaluated for any control u, by the quadratic performance index

$$J(p) = \lim_{t\to\infty} \frac{1}{t} \cdot E \int_0^t [(y^T(\tau,\bar{p})\ Q\ y(\tau,\bar{p})$$

$$+ u^T(\tau,p)\rho_c R_o\ u(\tau,p))]\ d\tau, \quad \text{scalar } \rho_c > 0 \qquad (3.7)$$

where Q, R_o are (kxk) and (mxm) symmetric positive definite matrices respectively.

For the case of deterministic system, the following modifications in the system description are in order.

i) Dw = 0, v = 0

ii) the initial condition, $x(0) = x_o$, $x_o x_o^T = \Sigma_o$ and the index V of reads

$$J(\bar{p}) = \int_0^\infty [y^T(t,\bar{p})\ Q\ y(t,\bar{p}) + u^T(t,\bar{p})\rho_c R_o\ u(t,\bar{p})]\ dt$$

If the state x(t) of the stochastic system (3.1) is estimated as a function of the measurements, we assume the 'state estimator' to be of the following structure

$$\hat{x}(t,\bar{p}) = A(\bar{p})\ \hat{x}(t,\bar{p}) + B(\bar{p})\ u(t,\bar{p}) + \hat{G}(\bar{p})\ \hat{z}(t,\bar{p}) \qquad (3.8)$$

where

$$\hat{z}(t,\bar{p}) = z(t,\bar{p}) - M(\bar{p})\ \hat{x}(t,\bar{p})$$

is called the 'measurement residual'. For a 'minimum variance' requirement, the estimator of (3.8) is the standard Kalman filter [13]. We refer the system presented in this section as the 'Basic System'.

3.2 Time Domain Stability Conditions Using Singular Value Decomposition:

Currently the 'Singular Value Decomposition (SVD)' has emerged as a useful tool in analyzing the stability robustness of multivariable control systems. Much of the published literature treats the robustness of feedback control systems in frequency domain. In this section, basic techniques are presented by which the concepts of SVD could be applied to the time domain analysis of 'stability robustness' of a linear time invariant system for the two cases of modeling errors: i) parameter variations and ii) truncated modes. Towards this direction, the matrix singular value properties are briefly reviewed.

3.2.1 Brief Review of Matrix Singular Value Properties:

Let F and E be two square matrices. Then

i) $\bar{\sigma}[F] \geq |\lambda[F]| \geq \underline{\sigma}[F]$

ii) If $\underline{\sigma}[F] = 0$ then F is singular

iii) If F^{-1} exists then $\bar{\sigma}[F] = 1/\underline{\sigma}[F^{-1}]$

where $\bar{\sigma}[\cdot]$ and $\underline{\sigma}[\cdot]$ denote the maximum and minimum singular values of the matrix $[\cdot]$ respectively and $\lambda[\cdot]$ denotes the eigenvalue of the matrix $[\cdot]$.

The main result which forms the basis for developing the stability for linear system conditions is stated in the following theorem[1]

Theorem 1: If F is nonsingular, then a sufficient condition for the matrix [F + E] to remain nonsingular is

$$\underline{\sigma}[F] > \bar{\sigma}[E]$$

The above theorem has an interesting implication. It suggests that if a given matrix F_1 is written as the sum of a nonsingular matrix F

and a perturbation on matrix E

i.e., $F_1 = F + E$ where F is nonsingular

then by use of theorem 1 one can get a condition for the nonsingularity of F_1.

Based on the above observation, in what follows, we present techniques to derive the stability conditions for a linear time invariant system having i) parameter variations and ii) truncated modes as the modeling errors.

3.2.2 Perturbations Due to Parameter Variations:

Let us consider a the linear time invariant system

$$\dot{x} = A(\bar{p})x + D(\bar{p})w, \quad x(0) = x_o \tag{3.9}$$

where the 'nominal' matrix $A(\bar{p})$ is asymptotically stable. In the presence of parameter variations in matrices A and D the perturbed system can be written as

$$\begin{bmatrix} \dot{x} \\ \dot{\Delta x} \end{bmatrix} = \begin{bmatrix} A & 0 \\ \Delta A & A+\Delta A \end{bmatrix} \begin{bmatrix} x \\ \Delta x \end{bmatrix} + \begin{bmatrix} D \\ \Delta D \end{bmatrix} w \tag{3.10}$$

where $\Delta[\cdot]$ denotes the total change in the matrix (or vector) $[\cdot]$ with no approximations about the size of the variation. Noting that the eigenvalues of the perturbed system are those of A and $A+\Delta A$, the stability of the perturbed system is determined by the stability of the perturbed matrix $A+\Delta A$. The straightforward application of theorem 1 then gives the following result.

Theorem A: The perturbed system due to parameter variations given by (3.10) is stable if

$$\underline{\sigma}[A] > \bar{\sigma}[\Delta A] \tag{3.11}$$

where A is an asymptotically stable matrix.

3.2.3 Perturbations Due to Truncated Modes:

Let us again consider the linear time invariant system

$$\dot{x} = A\,x + D\,w \qquad x(0) = x_o \tag{3.12}$$

where x is n x 1 vector and A is an asymptotically stable matrix. Let us suppose the perturbed system, which is a reduced order version of (3.12) (i.e., after mode truncation), is given by

$$\dot{x}_R = A_R\,x_R + D_R\,w \tag{3.13}$$

where x_R is an n_R x 1 vector where $n_R < n$ and A_R is not guaranteed to be stable. The aim is to derive the condition for A_R to be stable.

In order to do this, we augment the reduced order system of (3.13) by an articificial state vector x_d of dimension $(n - n_R)$ so that the composite vector $\underline{x}$ of dimension n is described by the differential equation

$$\dot{\underline{x}} = \begin{bmatrix} \dot{x}_R \\ \dot{x}_d \end{bmatrix} = \begin{bmatrix} A_R & 0 \\ 0 & -\epsilon \end{bmatrix} \begin{bmatrix} x_R \\ x_d \end{bmatrix} + \begin{bmatrix} D_R \\ 0 \end{bmatrix} w \tag{3.14a}$$

$$\dot{\underline{x}} = A_a\,\underline{x} + D_a\,w \qquad (A_a \text{ is an nxn matrix})$$

where ϵ is an $(n - n_R)$ x $(n - n_R)$ matrix having the form

$$\epsilon \triangleq \text{diag}\,[\epsilon_1, \epsilon_2, \ldots], \quad \epsilon_i > 0 \tag{3.14b}$$

with ε_i being arbitrarily small positive scalars so that $-\varepsilon$ is an asymptotically stable matrix. Thus the system of (3.14) is asymptotically stable only if A_R is asymptotically stable.

Treating A_a as the perturbed system matrix and writing the asymptotically stable matrix A as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad , \; A_{11} \text{ is an } n_R \times n_R \text{ matrix} \tag{3.15}$$

we can write the perturbation matrix E as

$$E = A_a - A = \begin{bmatrix} A_R & 0 \\ 0 & -\varepsilon \end{bmatrix} - \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_R - A_{11} & -A_{12} \\ -A_{12} & -\varepsilon - A_{22} \end{bmatrix} \tag{3.16}$$

then

$$A_a = A + E \tag{3.17}$$

To guarantee the stability of the perturbed matrix A_a one can apply theorem 1 and get the following condition

Theorem B1 The perturbed system matrix A_a of (3.17) is stable if

$$\underline{\sigma}[A] > \bar{\sigma}[E] \tag{3.18}$$

where E is given by (3.16).

Since ε is a matrix with arbitrarily small norm, it is possible to the E matrix, in the limit as $\frac{\varepsilon_{ij}}{A_{22ij}} \to 0$ (when $A_{22ij} \neq 0$)

$$E_m = \begin{bmatrix} A_R - A_{11} & -A_{12} \\ -A_{21} & -A_{22} \end{bmatrix} \tag{3.19}$$

Thus, theorem B1 can be restated as

Theorem B2: The perturbed system matrix A_a of (3.17) is stable if

$$\underline{\sigma}[A] > \overline{\sigma}[E_m] \tag{3.20}$$

where E_m is given by (3.19).

Theorems A and B form the basis in extending the stability conditions to the case of LQG regulators for various cases of modeling errors involving parameter variations and model/controller truncation. These are discussed in the following sections.

3.3 'Stability Robustness' Analysis of LQG Regulators for Various Cases of Perturbation

In this section, the results of section 3.2 are extended to the case of LQG regulators having the system description of section 3.1. Perturbation matrices are formed for various combinations of the two modeling errors i) parameter variations and ii) model/controller truncation, and then the results of SVD of Section 3.2 are applied to arrive at conditions for the stability of the closed loop system.

Case 1: Parameter variations alone, No model/controller truncation:

For this case, the following assumptions are made with respect to the model described by equations (3.1).

Assumption 1: The matrix pairs $[A(\bar{p}), B(\bar{p})]$ and $[A(\bar{p}), D(\bar{p})]$ are completely controllable and the pairs $[A(\bar{p}), C(\bar{p})]$ and $[A(\bar{p}), M(\bar{p})]$ are completely observable.

Since there is no model/controller truncation the full order optiomal control for nominal values of the parameters is given by

$$u = G\hat{x} = -\frac{1}{\rho_c} R_o^{-1} B^T K \hat{x} \tag{3.21a}$$

where

$$\dot{\hat{x}} = A\hat{x} + Bu + \hat{G}(z - M\hat{x}), \quad \hat{x}(0) = 0 \tag{3.21b}$$

$$= (A + BG - \hat{G}M)\hat{x} + \hat{G}z \tag{3.21c}$$

$$\hat{G} = \frac{1}{\rho_e} P M^T V_o^{-1} \tag{3.21c}$$

and P and K satisfy the algebraic matrix Riccati equations

$$KA + A^TK - KB \frac{R_o^{-1}}{\rho_c} B^TK + C^TQC = 0 \tag{3.21d}$$

$$PA^T + AP - PM^T \frac{V_o^{-1}}{\rho_e} MP + DWD^T = 0 \tag{3.21e}$$

The nominal closed loop system is given by

$$\begin{bmatrix} \dot{x} \\ \dot{\hat{x}} \end{bmatrix} = \begin{bmatrix} A & BG \\ \hat{G}M & \hat{A}_c \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \end{bmatrix} + \begin{bmatrix} D & 0 \\ 0 & \hat{G} \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix} \tag{3.22a}$$

$$\begin{bmatrix} y \\ u \end{bmatrix} = \begin{bmatrix} C & 0 \\ 0 & G \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \end{bmatrix} \tag{3.22b}$$

where $\hat{A}_c = A + BG - \hat{G}M$ and closed loop system is asymptotically stable.

We are now interested in examining the stability robustness of the closed loop system in the presence of parameter variations alone. Let ΔA, ΔB, ΔC, ΔM and ΔD be the perturbation in the system matrices A, B, C M and D respectively. Then the perturbed system can be written as (Note: the filter parameters $\hat{A}_c$, $\hat{G}$ and the control gain G are not subjected to variations because they are specified by the designer as a function of the nominal values of the parameters.)

$$\begin{bmatrix} \dot{x} \\ \dot{\hat{x}} \\ \dot{\Delta x} \\ \dot{\hat{\Delta x}} \end{bmatrix} = \begin{bmatrix} A & BG & 0 & 0 \\ \hat{G}M & \hat{A}_c & 0 & 0 \\ \Delta A & \Delta BG & A+\Delta A & (B+\Delta B)G \\ \hat{G}\Delta M & 0 & \hat{G}(M+\Delta M) & \hat{A}_c \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \\ \Delta x \\ \hat{\Delta x} \end{bmatrix} + \begin{bmatrix} D & 0 \\ 0 & \hat{G} \\ \Delta D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix} \tag{3.23}$$

Then by the application of theorem A of section 3.2, we obtain the following design observation.

Design Observation 1: The perturbed LQG regulator system is stable if

$$\underline{\sigma}[A_{cL}] > \bar{\sigma}[E_{p.v}] \tag{3.24}$$

where

$$A_{cL} = \begin{bmatrix} A & BG \\ \hat{G}M & \hat{A}_c \end{bmatrix} \quad \text{and,} \quad E_{p.v} = \begin{bmatrix} \Delta A & \Delta BG \\ \hat{G}\Delta M & 0 \end{bmatrix} \tag{3.25}$$

Case 2: Perturbations due to Controller Truncation Alone, No Model Reduction and No Parameter Variations

In this case, we assume the full order controller is given by equations (3.21). To meet on-line controller software constraints let us suppose it is intended to retain n_c of the n controller states ($n_c < n$) and truncate the remaining ($n - n_c$) controller states. The question of which n_c controller states are to be retained is in itself an important area of research and is not addressed in this report. Let the reduced order controller be described by

$$u = \bar{G}_R \bar{x}_R \tag{3.26a}$$

$$\dot{\bar{x}}_R = \bar{A}_R \bar{x}_R + \bar{G}_R z, \qquad \bar{x}_R \in R^{n_c} \tag{3.26b}$$

where $\bar{x}_R$ may be either a direct truncation of $\hat{x}$ [e.g. $\bar{G}_R$ is obtained by

deleting the last $\bar{n}_c (= n - n_c)$ columns of G, $\bar{\bar{G}}_R$ is obtained by deleting the last $\bar{n}_c$ rows of $\hat{G}$ and $\bar{\bar{A}}_R$ is obtained by deleting the corresponding rows and columns of $\hat{A}_c$ of (3.25)] or it could be obtained as a partial realization of $\hat{x}$ involving some controller reduction technique.

Then the perturbed closed loop system matrix is given by

$$A_{cL_2} = \begin{bmatrix} A & B\bar{\bar{G}}_R \\ \bar{\bar{G}}_R M & \bar{\bar{A}}_R \end{bmatrix} \quad \text{whose stability is to be} \tag{3.27}$$

established. (Here A_{cL_2} is an $(n + n_c) \times (n + n_c)$ matrix).

Denoting the full order controller matrices as

$$G = [G_R \quad G_T] \tag{3.28a}$$

$$\hat{G} = \begin{bmatrix} \hat{G}_R \\ \hat{G}_T \end{bmatrix} \quad ; \quad \hat{A}_c = \begin{bmatrix} \hat{A}_R & \hat{A}_{RT} \\ \hat{A}_{TR} & \hat{A}_T \end{bmatrix} \tag{3.28b}$$

where the first partition is of dimension n_c, the nominally asymptotically stable closed loop system matrix A_{cL} of (3.25) can be written as

$$A_{cL} = \begin{bmatrix} A & BG_R & BG_T \\ \hat{G}_R M & \hat{A}_R & \hat{A}_{RT} \\ \hat{G}_T M & \hat{A}_T R & \hat{A}_T \end{bmatrix} \quad , \text{ where } A_{cL} \text{ is a } 2n \times 2n \text{ matrix.} \tag{3.29}$$

Employing the technique given for truncated mode case in section (3.23), the condition for the stability of the perturbed closed loop system matrix A_{cL_2} can be stated as follows:

Design Observation 2: The perturbed system matrix A_{cL_2} of (3.27) is stable if

$$\underline{\sigma}[A_{cL}] > \bar{\sigma}[E_{T.M}] \tag{3.30}$$

where

$$E_{T.M.} = \begin{bmatrix} 0 & B(\bar{G}_R - G_R) & -BG_T \\ (\bar{G}_R - \hat{G}_R)M & \bar{A}_R - \hat{A}_R & -\hat{A}_{RT} \\ -\hat{G}_T M & -\hat{A}_{TR} & -\hat{A}_T \end{bmatrix} \tag{3.31}$$

Note: In a controller reduction scheme where $\bar{G}_R \equiv G_R$, $\hat{G}_R \equiv G_R$ and $\bar{A}_R \equiv \hat{A}_R$ the matrix $E_{T.M}$ simplifies to

$$E_{T.M} = \begin{bmatrix} & & -BG_T \\ & \bigcirc & -\hat{A}_{RT} \\ -\hat{G}_T M & -\hat{A}_{TR} & -\hat{A}_T \end{bmatrix} \tag{3.32}$$

Case 3: Combined Case of Parameter Variations and Controller Truncation (No Model Reduction):

For this case, the perturbed system matrix is described by

$$A_{cL_3} = \begin{bmatrix} A + \Delta A & (B + \Delta B)\bar{G}_R \\ \bar{G}_R(M + \Delta M) & \bar{A}_R \end{bmatrix} \tag{3.33}$$

and thus it can be deduced that the condition for stability of A_{cL_3} is as follows:

Design Observation 3: The perturbed system matrix A_{cL_3} is stable if

$$\underline{\sigma}[A_{cL}] > \bar{\sigma}[E_{c.c}] \tag{3.34}$$

where

$$E_{c.c} = \begin{bmatrix} \Delta A & B(G_R - \bar{G}_R) + \Delta B\bar{G}_R & -BG_T \\ (\bar{G}_R - \hat{G}_R)M + \bar{G}_R\Delta M & \bar{A}_R - \hat{A}_R & -\hat{A}_{RT} \\ -\hat{G}_T M & -\hat{A}_{TR} & -\hat{A}_T \end{bmatrix} \tag{3.35}$$

Case 4: Model Reduction Alone, No Controller Reduction and No Parameter Variations

For this case, we treat the model given by (3.1) as the evaluation model. We assume the order of the model, n to be too high for control u to be determined and that there is a control design model of dimension $n_R < n$ given by

$$\begin{aligned} \dot{\bar{x}}_R &= \bar{A}_R \bar{x}_R + \bar{B}_R u + \bar{D}_R w \quad , \qquad \bar{x}_R \in R^{n_R} \\ \bar{y}_R &= \bar{C}_R \bar{x}_R \\ \bar{z}_R &= \bar{M}_R \bar{x}_R + v \end{aligned} \tag{3.36}$$

where the above control design model is obtained either by a direct truncation of the full order model given by

$$\dot{x} = \begin{bmatrix} \dot{x}_R \\ \dot{x}_T \end{bmatrix} = \begin{bmatrix} A_R & A_{RT} \\ A_{TR} & A_T \end{bmatrix} \begin{bmatrix} x_R \\ x_T \end{bmatrix} + \begin{bmatrix} B_R \\ B_T \end{bmatrix} u + \begin{bmatrix} D_R \\ D_T \end{bmatrix} w \tag{3.37a}$$

$$y = \begin{bmatrix} C_R & C_T \end{bmatrix} \begin{bmatrix} x_R \\ x_T \end{bmatrix} \tag{3.37b}$$

$$z = \begin{bmatrix} M_R & M_T \end{bmatrix} \begin{bmatrix} x_R \\ x_T \end{bmatrix} \quad ; \quad x_R \in R^{n_R} \tag{3.37c}$$

or by a partial realization of x involving some model reduction technique.

Let the optimal control for the control design model of (3.36) be obtained by minimizing the performance index

$$J_R = \lim_{t \to \infty} \frac{1}{t} \int_0^t (\bar{y}_R^T \, Q \, \bar{y}_R + u^T \rho_c \, R_o \, u) \, dt \tag{3.38}$$

Assuming the pairs

$$(\bar{A}_R , \bar{B}_R), \quad (\bar{A}_R, \bar{D}_R) \quad \text{to be c. controllable}$$

and $(\bar{A}_R , \bar{C}_R), \quad (\bar{A}_R, \bar{M}_R)$ to be c. observable (3.39

the full order optimal control for the reduced order model is given by

$$u = \bar{G}_R \bar{\bar{x}}_R \tag{3.40}$$

$$\dot{\bar{\bar{x}}}_R = \bar{A}_R \bar{\bar{x}}_R + \bar{B}_R u + \bar{\bar{G}}_R (z - \bar{M}_R \bar{\bar{x}}_R) \tag{3.40}$$

$$= \bar{\bar{A}}_R \bar{\bar{x}}_R + \bar{\bar{G}}_R z$$

where

$$\bar{G}_R = - \frac{1}{\rho_c} R_o^{-1} \bar{B}_R^T K_R \tag{3.40}$$

$$\bar{\bar{G}}_R = \frac{1}{\rho_e} P_R \bar{M}_R^T V_o^{-1} \tag{3.40}$$

$$K_R \bar{A}_R + \bar{A}_R^T K_R - K_R \bar{B}_R \frac{R_o^{-1}}{\rho_c} \bar{B}_R^T K_R + \bar{C}_R^T Q \bar{C}_R = 0 \tag{3.40}$$

$$P_R \bar{A}_R^T + \bar{A}_R P_R - P_R \bar{M}_R^T \frac{V_o^{-1}}{\rho_e} \bar{M}_R P_R + \bar{D}_R W \bar{D}_R^T = 0 \tag{3.40}$$

The closed loop system matrix for the control design model is given by

$$\bar{A}_{cL} = \begin{bmatrix} \bar{A}_R & \bar{B}_R \bar{G}_R \\ \bar{\bar{G}}_R \bar{M}_R & \bar{\bar{A}}_R \end{bmatrix} \tag{3.41}$$

and is asymptotically stable.

The closed loop system for the evaluation model is obtained by forcing the evaluation model with the controller of the control design model. Thus, we have

$$\begin{bmatrix} \dot{x}_R \\ \dot{\bar{x}}_R \\ \dot{x}_T \end{bmatrix} = \begin{bmatrix} A_R & B_R G_R & A_{RT} \\ \bar{G}_R M_R & \bar{A}_R & \bar{G}_R M_T \\ A_{TR} & B_T \bar{G}_R & A_T \end{bmatrix} \begin{bmatrix} x_R \\ \bar{x}_R \\ x_T \end{bmatrix} + \begin{bmatrix} D_R & 0 \\ 0 & \bar{G}_R \\ D_T & 0 \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix} \quad (3.4$$

$$\begin{bmatrix} y \\ u \end{bmatrix} = \begin{bmatrix} C_R & 0 & C_T \\ 0 & \bar{G}_R & 0 \end{bmatrix} \begin{bmatrix} x_R \\ \bar{x}_R \\ x_T \end{bmatrix}$$

and

$$J = \lim_{t\to\infty} \frac{1}{t} \int_0^t [y^T \quad u^T] \begin{bmatrix} Q & 0 \\ 0 & \rho_c R_o \end{bmatrix} \begin{bmatrix} y \\ u \end{bmatrix} dt \quad (3.43$$

The stability of the closed loop system matrix

$$A_{cL_4} = \begin{bmatrix} A_R & B_R \bar{G}_R & A_{RT} \\ \bar{G}_R M_R & \bar{A}_R & \bar{G}_R M_T \\ A_{TR} & B_T \bar{G}_R & A_T \end{bmatrix}$$ is to be established. (3.44

At this juncture we assume the matrix A_T of (3.37) to be an asymptotically stable matrix (which is a reasonable assumption for large space structure models). In order to derive the condition for stability of the closed loop system matrix A_{cL_4}, we employ the technique given in Section (3.2.6) for truncated mode case. Accordingly we write A_{cL_4} of (3.44) as

$$A_{cL_4} = \begin{bmatrix} \bar{A}_R & \bar{B}_R \bar{G}_R & 0 \\ \bar{G}_R \bar{M}_R & \bar{A}_R & 0 \\ 0 & 0 & A_T \end{bmatrix} + \begin{bmatrix} A_R - \bar{A}_R & (B_R - \bar{B}_R)\bar{G}_R & A_{RT} \\ \bar{G}_R (M_R - \bar{M}_R) & 0 & \bar{G}_R M_T \\ A_{TR} & B_T \bar{G}_R & 0 \end{bmatrix} \quad (3.45$$

$A_{cL_4} = A_{SC} + E_{F.L}$ where A_{SC} is seen to be an asymptotically stable matrix. Thus, we have

Design Observation 4: The closed loop system matrix A_{cL_4} is stable if

$$\underline{\sigma}[A_{SC}] > \bar{\sigma}[E_{F.L}] \tag{3.4}$$

where A_{SC} & $E_{F.L}$ are given by (3.45).

The motivation for writing A_{cL_4} as given by (3.45) is that the matrices A_{SC} and $E_{F.L}$ can be easily constructed from the known matrices of the control design model and the open loop evaluation model. It is to be noted that the matrix $E_{F.L}$ containing basically the terms that cause 'spillover'.

Case 5: Model Reduction and Parameter Variations: No controller Reduction

In this case the perturbation matrix takes the following form

$$E_{F.P.V.} = \begin{bmatrix} A_R-\bar{A}_R+\Delta A_R & (B_R-\bar{B}_R)\bar{G}_R+\Delta B_R\bar{G}_R & A_{RT}+\Delta A_{RT} \\ \bar{G}_R(M_R-\bar{M}_R)+\bar{G}_R\Delta M_R & 0 & \bar{G}_R(M_T+\Delta M_T) \\ A_{TR}+\Delta A_{TR} & (B_T+\Delta B_T)\bar{G}_R & \Delta A_T \end{bmatrix} \tag{3.}$$

and the condition for stability is as follows:

Design Observation 5: The closed loop system matrix $A_{cL_4} + \Delta A_{cL_4}$ is stable if

$$\underline{\sigma}[A_{SC}] > \bar{\sigma}[E_{F.P.V}] \text{ and } \underline{\sigma}[A_{SC}] > \bar{\sigma}\,[E_{F.L}] \tag{3.}$$

where A_{SC} is given by (3.45) and $E_{F.P.V}$ is given by (3.47) and

$$A_{SC} + E_{F.P.V} = A_{cL_4} + \Delta A_{cL_4}$$

$$= \begin{bmatrix} A_R+\Delta A_R & (B_R+\Delta B_R)\bar{G}_R & A_{RT}+\Delta A_{RT} \\ \bar{\bar{G}}_R(M_R+\Delta M_R) & \bar{\bar{A}}_R & \bar{\bar{G}}_R(M_T+\Delta M_T) \\ A_{TR}+\Delta A_{TR} & (B_T+\Delta B_T)\bar{G}_R & A_T+\Delta A_T \end{bmatrix} \quad (3.49)$$

Case 6: Model Reduction and Controller Reduction, No Parameter Variations

For this case, we assume the controller of dimension n_R given by equations (3.40) is reduced to $n_c (n_c < n_R)$ be deleting the last $(n_R - n_c)$ rows of $\bar{\bar{G}}_R$, the last $(n_R - n_c)$ columns of $\bar{G}_R$ and the corresponding rows and columns of $\bar{\bar{A}}_R$. Accordingly we write

$$\bar{x}_R = \begin{bmatrix} \bar{x}_{R_1} \\ \bar{x}_{R_2} \end{bmatrix} \quad (3.50a)$$

$$\bar{G}_R = [\bar{G}_{R_1} , \bar{G}_{R_2}] \quad (3.50b)$$

$$\bar{\bar{G}}_R = \begin{bmatrix} \bar{\bar{G}}_{R_1} \\ \bar{\bar{G}}_{R_2} \end{bmatrix} \quad \text{and} \quad \bar{\bar{A}}_R = \begin{bmatrix} \bar{\bar{A}}_{R_{11}} & \bar{\bar{A}}_{R_{12}} \\ \bar{\bar{A}}_{R_{21}} & \bar{\bar{A}}_{R_{22}} \end{bmatrix} \quad (3.50c)$$

where the first partition is of dimension n_c. Then the closed loop system with reduced order controller is given by

$$\begin{bmatrix} \dot{x}_R \\ \dot{\bar{x}}_{R_1} \\ \dot{x}_T \end{bmatrix} = \begin{bmatrix} A_R & B_R\bar{G}_{R_1} & A_{RT} \\ \bar{G}_{R_1}M_R & \bar{A}_{R_{11}} & \bar{G}_{R_1}M_T \\ A_{TR} & B_T\bar{G}_{R_1} & A_T \end{bmatrix} \begin{bmatrix} x_R \\ \bar{x}_{R_1} \\ x_T \end{bmatrix} + \begin{bmatrix} D_R & 0 \\ 0 & \hat{G}_{R_1} \\ D_T & 0 \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix} \quad (3.$$

$$\begin{bmatrix} y \\ u \end{bmatrix} = \begin{bmatrix} C_R & 0 & C_T \\ 0 & \bar{G}_{R_1} & 0 \end{bmatrix} \begin{bmatrix} x_R \\ \bar{x}_{R_1} \\ x_T \end{bmatrix}$$

Then employing the technique of section (3.23), the perturbed system matrix [the closed loop system matrix of (3.51), denoted by A_{cL_6}, can be written as

$$\begin{bmatrix} A_R & B_R\bar{G}_{R_1} & 0 & A_{R_T} \\ \bar{G}_{R_1}M_R & \bar{A}_{R_{11}} & 0 & \bar{G}_{R_1}M_T \\ 0 & 0 & -\varepsilon & 0 \\ A_{TR} & B_T\bar{G}_{R_1} & 0 & A_T \end{bmatrix} = \begin{bmatrix} \bar{A}_R & \bar{B}_R\bar{G}_{R_1} & \bar{B}_R\bar{G}_{R_2} & 0 \\ \bar{G}_{R_1}\bar{M}_R & \bar{A}_{R_{11}} & \bar{A}_{R_{22}} & 0 \\ \bar{G}_{R_2}\bar{M}_R & \bar{A}_{R_{21}} & \bar{A}_{R_{22}} & 0 \\ 0 & 0 & 0 & A_T \end{bmatrix}$$

$$+ \begin{bmatrix} A_R-\bar{A}_R & (B_R-\bar{B}_R)\bar{G}_{R_1} & -\bar{B}_R\bar{G}_{R_2} & A_{RT} \\ \bar{G}_{R_1}(M_R-\bar{M}_R) & 0 & -\bar{A}_{R_{12}} & \bar{G}_{R_1}M_T \\ -\bar{G}_{R_2} & -\bar{A}_{R_{21}} & -\bar{A}_{R_{22}}-\varepsilon & 0 \\ A_{TR} & B_T\bar{G}_{R_1} & 0 & 0 \end{bmatrix} \quad ($$

i.e. $A_{cL_6} = A_{SC} + E'_{P.L}$ (3.54)

Applying theorem B2 of Section (3.23), the condition for stability of the perturbed system matrix A_{cL_6} is given by the following:

Design Observation 6: The closed loop system matrix A_{cL_6} is stable if

$$\underline{\sigma}[A_{SC}] > \overline{\sigma}[E_{P.L}] \tag{3.55}$$

where $E_{P.L}$ is the matrix $E'_{P.L}$ of (3,54) with $\varepsilon = 0$.

We finally have the combined case of model reduction, controller reduction and parameter variations.

Case 7: Model Reduction, Controller Reduction and Parameter Variations

In this case, the perturbation matrix takes the following form

$$E_{P.P.V} = \begin{bmatrix} A_R-\bar{A}_R+\Delta A_R & (B_R-\bar{B}_R)\bar{G}_R+\Delta B_R\bar{G}_{R_1} & -\bar{B}_R\bar{G}_{R_2} & A_{RT}+\Delta A_{RT} \\ \bar{G}_{R_1}(M_R-\bar{M}_R)+\bar{G}_{R_1}\Delta M_R & 0 & -\bar{A}_{R_{12}} & \bar{G}_{R_1}(M_T+\Delta M_T) \\ -\bar{G}_{R_2}\bar{M}_R & -\bar{A}_{R_{21}} & -\bar{A}_{R_{22}} & 0 \\ A_{TR}+\Delta A_{TR} & (B_T+\Delta B_T)\bar{G}_{R_1} & 0 & A_T+\Delta A_T \end{bmatrix} \tag{3.56}$$

and the condition for stability is as follows:

Design Observation 7: The closed loop system matrix $A_{cL_6} + \Delta A_{cL_6}$ is stable if

$$\underline{\sigma}[A_{s.c}] > \overline{\sigma}[E_{P.P.V}] \text{ and } \underline{\sigma}[A_{SC}] > \overline{\sigma}\ [E_{P.L}] \tag{3.57}$$

where $E_{P.P.V.}$ is given by (3.56).

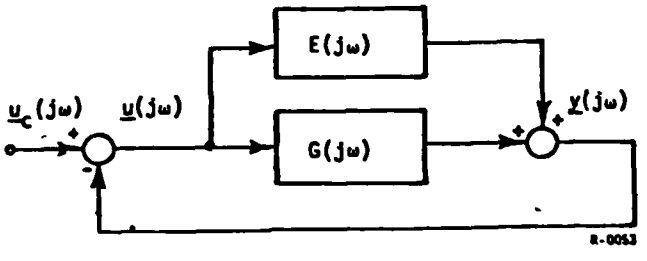
3.4 Discussion of the Design Observations:

Some discussion about the implications of these design observations is now in order. Firstly, it may be noted that the proposed stability conditions are similar, conceptually, to the frequency domain results reported in Ref.[1]. However there are also some interesting differences between these two (frequency domain and time domain) versions. Some preliminary observations are presented in the following sections. Secondly, these design observations are useful in many ways in both the analysis and synthesis of robust controllers. These are discussed in later sections.

3.4.1 Comparison and Contrast Between Frequency Domain Analysis and the Proposed Time Domain Analysis:

In refs [1-3] the following unity feedback system with additive perturbations is considered.

Fig. 1 Additively Perturbed Feedback System



For this system the stability conditions are reported as

$$\underline{\sigma}[I + G(j\omega)] > \overline{\sigma}[E(j\omega)] \tag{3.58}$$

where $[I + G(j\omega)]$ is the 'Return difference matrix' and $E[(j\omega)]$ is the 'perturbation matrix'. The above two matrices are seen to be functions of frequency. In the case of time invariant linear systems of (3.1), the frequency dependency is due to the presence of the state transition matrix $\Phi(j\omega) = [j\omega I - A]^{-1}$ since all the other matrices are constant matrices. In the proposed time domain analysis the stability conditions are in terms of the closed loop system matrix and a perturbation matrix which are seen to be constant matrices. In the case of frequency domain, from relation (3.58), the perturbed system matrix is $[I + G(j\omega) + E(j\omega)]$. Indeed its counterpart in time domain would be a function of $e^{\bar{A}t}$ where the perturbed matrix $\bar{A}$ can be written as

$$\bar{A} = A_s + E \tag{3.59}$$

where A_s is nominally asymptotically stable matrix and E is an additive perturbation matrix. In fact the stability conditions can then be in terms of

$$\underline{\sigma}[A_s t] \text{ and } \overline{\sigma}[Et] \tag{3.60}$$

which are seen to be functions of time. However one realizes from the above form that the necessary information one needs is $\underline{\sigma}[A_s]$ and $\overline{\sigma}[E]$ and this is the form taken by the proposed stability conditions. Thus the main difference between the frequency domain version and the time domain version is that,

- In the stability conditions of frequency domain, a singular value

vs. frequency plot is to be drawn.

· In the stability conditions of time domain, no time dependency is present. Only two scalars are to be compared.

ii) In the case of frequency domain results, the perturbations are mainly viewed in terms of 'gain' and 'phase' changes[4,5]. In the proposed time domain analysis the perturbations are viewed as 'system parameter variations' and 'system model/controller order with constant, fixed, gains. It may be noted that in the time domain treatment the nominally stable matrix (left hand side matrix of the stability condition) and the perturbation matrix (right hand side matrix of the stability condition) are both functions of the constant controller gains.

iii) In the frequency domain treatment, the requirement of square matrices necessitates the assumption that the number of control inputs be equal to the number of outputs, even though this assumption can be satisfied most of the time by appropriate selection of the break point of the loop. However in the present time domain analysis no such assumption is needed.

iv) In the proposed time domain analysis, the model-error information (i.e. parameter variation, mode truncation) is explicitly incorporated into the formulation. Also in the proposed stability conditions, provision is made for considering different reduced order models for control design purposes (reflected by the presence of $\bar{A}_{R_1}, \bar{B}_{R_1}$ matrices which could be different from A_{R_1}, B_{R_1} ... matrices). Just as in frequency domain different control design methods could be compared using the singular value plots of the 'Return-difference matrices', this provision helps to compare to different reduced order models (used for control design)

from 'stability robustness' point of view.

These are some of the preliminary observations made with to the frequency domain and time domain approaches for 'stability robustness'. Evidently further in roads have to be made in the investigation of this relationship and this is suggested as a future-research topic. In following section, the usefulness of the proposed design observations is briefly discussed.

3.4.2 Usefulness of the Design Observations

The proposed design observations are helpful in many ways.

Firstly, if a nominally stable closed loop system matrix is specified (i.e. either A_{cL} or $A_{s.c}$, depending on the specific case as discussed in section (3.3)), then one can use these tests to determine the tolerablw perturbations in the system matrices A, B and M and the order of the model/controller before the closed loop system becomes unstable.

Conversely, given the perturbations ΔA, ΔB and ΔM and the order of the model/controller, one may determine the controller gains to achieve stability robustness. In a practical situation the entries of the matrices ΔA, ΔB and ΔM could be some reasonable absolute increments expected in the entries A, B and M respectively or in the case when the functional dependence of matrices A, B and M is known in terms of an uncertain parameter vector p with a statistical model, the entries of ΔA, ΔB and ΔM get determined accordingly.

As indicated in the previous section, these tests can be used to compare different model reduction and control design schemes from stability

robustness point of view.

Finally these tests can find applications in spillover reduction problems and sensor/actuator location problems.

3.5 Measure of Stability Robustness:

It may be noted from the stability conditions derived in Section 3.3, that these conditions may be satisfied by a variety of controller gains for a given perturbations. In order to compare different models/controllers from stability robustness point of view it is clear that there is a need for developing a relative measure of stability robustness. Once this measure is developed, it may also be used as a performance index to optimize the parameters of the model/controller. In what follows we define a measure of 'stability robustness' based on the results of Section 3.3.

Let us define the 'stability robustness index' as

$$\beta_{S.R.} \triangleq \frac{\underline{\sigma}[A_{st}] - \overline{\sigma}[E]}{\underline{\sigma}[A_{st}]} \qquad (3.6$$

where $\underline{\sigma}[A_{st}] \geq \overline{\sigma}[E]$, A_{st} is the asymptotically stable matrix present on the L.H.S. of the condition which could be either $A_{s.c}$ or $A_{C.L.}$ depending on the case considered and E is the perturbation matrix present on the R.H.S. of the condition and takes the appropriate form described in the stability condition of Section 3.3 depending on the case considered.

From this definition, it is clear that the range of β is given by $0 \leq \beta \leq 1$. Thus, of all the model/controllers which satisfy the stability condition, the system with one extreme $\beta_{S.R.} = 1$ has the

maximum stability robustness and with the other extreme $\beta_{S.R.} = 0$ has zero stability robustness.

The measure above measure can be used to compare different model/controller combinations with given perturbations. In fact, in principle, for control design purposes one can even pose the following optimization problem:

Maximize $[\beta_{S.R.}]$

with respect to design parameters ρ_c and ρ_e of (3.3 and 3.7).

Even though an analytical solution to the above optimization problem is a formidable task, it may be possible to solve it by an iterative numerical procedure. However, it is not our intention to maximize $\beta_{S.R}$ because we are not only interested in 'stability robustness' but also in 'performance robustness'. As stated in Section (2.0) of this report the objective is to obtain a reasonable trade-off between stability and performance. Within this perspective , in what follows, we address the aspect of 'Performance Robustness'.

IV. Time Domain 'Performance Robustness' Analysis of LQG Regulators:

4.1 'Performance Robustness' for LQG Regulators:

It is noted that performance robustness, in general, is meant to maintain satisfactory level of performance in the presence of modeling errors and that performance robustness analysis assumes stability of the closed loop system. This perspective holds good even for the case of LQG regulators having the system description given in Section 3.1. Accordingly, in the following performance robustness studies we

assume the control design is such that the stability conditions as developed in Section 3.3 are satisfied (depending on the case under consideration (Section 3.3) the appropriate stability condition is assumed satisfied). For ready applicability, we will document the closed loop system equations for the most general case, namely Case 7 (with parameter variations and model/controller truncation). The resulting system equations are given by

$$\begin{bmatrix} \dot{x} \\ \dot{\bar{x}}_{R_1} \\ \Delta\dot{x} \\ \Delta\dot{\bar{x}}_{R_1} \end{bmatrix} = \begin{bmatrix} A & \bar{B}\bar{G}_{R_1} & 0 & 0 \\ \bar{G}_{R_1}M & \bar{A}_{R_{11}} & 0 & 0 \\ \nabla A & \Delta B\bar{G}_{R_1} & A+\Delta A & (B+\Delta B)\bar{G}_{R_1} \\ \bar{G}_{R_1}\Delta M & 0 & \bar{G}_{R_1}(M+\Delta M) & \bar{A}_{R_{11}} \end{bmatrix} \begin{bmatrix} x \\ \bar{x}_{R_1} \\ \Delta x \\ \Delta\bar{x}_{R_1} \end{bmatrix} + \begin{bmatrix} D & 0 \\ 0 & \bar{G}_{R_1} \\ \Delta D & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} w \\ v \end{bmatrix}$$

$$\begin{bmatrix} y \\ \Delta y \\ u \\ \Delta u \end{bmatrix} = \begin{bmatrix} C & 0 & 0 & 0 \\ \Delta C & 0 & (C+\Delta C) & 0 \\ 0 & \bar{G}_{R_1} & 0 & 0 \\ 0 & 0 & 0 & \bar{G}_{R_1} \end{bmatrix} \cdot \begin{bmatrix} x \\ \bar{x}_{R_1} \\ \Delta x \\ \Delta\bar{x}_{R_1} \end{bmatrix}$$

where $\bar{x}_R = \begin{bmatrix} \bar{x}_{R_1} \\ \bar{x}_{R_2} \end{bmatrix}$, $\bar{x}_{R_1} \in R^{n_c}$ and $\Delta[\cdot]$ is the total change in the matrix (or vector) $[\cdot]$ with no approximations about the size of the variation. The above system can be written as

$$\dot{x}_f = A_f x_f + D_f w_f \quad , \quad x_f \epsilon R^{2(n+n_c)}$$

$$y_f = C_f x_f$$

$$\text{where } x_f^T = [x^T \; \bar{x}_{R_1}^T \quad (\Delta x)^T \quad \Delta\bar{x}_{R_1}] \; , \quad w_f^T = [w^T \quad v^T] \tag{4.2}$$

$$y_f^T = [y^T \; (\Delta y)^T \quad u^T \quad (\Delta u)^T], \qquad E[w_f(t) w_f^T(\tau)] = \begin{bmatrix} W & 0 \\ 0 & \rho_e V_o \end{bmatrix} = W_f$$

and A_f, D_f and C_f are the corresponding matrices in equations (4.1).

The performance of the system can be evaluated by the index

$$J_{c.l.\ loop} = \lim_{t\to\infty} \frac{1}{t} \int_o^t [y(\bar{p})+\Delta y]^T Q[y(\bar{p})+\Delta y] + [u(\bar{p})+\Delta u]^T \rho_c R_o [u(\bar{p})+\Delta u] dt$$

$$= \lim_{t\to\infty} \frac{1}{t} \int_o^t (y_f^T Q_f y_f) dt \quad \text{where } Q_f = \begin{bmatrix} Q & Q & 0 & 0 \\ Q & Q & 0 & 0 \\ 0 & 0 & \rho_c R_o & \rho_c R_o \\ 0 & 0 & \rho_c R_o & \rho_c R_o \end{bmatrix} \tag{4.3}$$

The cost $J_{a.loop}$ can be computed as follows:

$$J_{c.l.loop} = \text{trace } P_f C_f^T Q_f C_f \tag{4.4a}$$

where P_f satisfies the Liapunov equation

$$P_f A_f^T + A_f P_f + D_f W_f D_f^T = 0 \tag{4.4b}$$

Because of the special structure of $A_f = \begin{bmatrix} A_{f_{11}} & 0 \\ A_{f_{21}} & A_{f_{22}} \end{bmatrix}$, the (4.5)

$2(n+n_c)^{th}$ order Liapunov equation given by (4.4b) can be dismembered into 3 Liapunov equations each of order $(n+n_c)$ [Note: $\text{Max}(n+n_c) = 2n$] as follows:

$$P_{f_{11}} A_{f_{11}}^T + A_{f_{11}} P_{f_{11}} + D_{f_{11}} W D_{f_{11}}^T = 0$$

$$P_{f_{12}} A_{f_{12}}^T + A_{f_{11}} P_{f_{12}} + P_{f_{11}} A_{f_{21}}^T + D_{f_{11}} W D_{f_{21}}^T = 0 \qquad (4.$$

$$P_{f_{22}} A_{f_{22}}^T + A_{f_{22}} P_{f_{22}} + P_{f_{12}}^T A_{f_{21}}^T + A_{f_{21}} P_{f_{12}} + D_{f_{21}} W D_{f_{21}}^T = 0$$

where

$$P_f = \begin{bmatrix} P_{f_{11}} & P_{f_{12}} \\ P_{f_{12}}^T & P_{f_{22}} \end{bmatrix} \quad ; \quad D_f = \begin{bmatrix} D_{f_{11}} \\ D_{f_{21}} \end{bmatrix}$$

and the first partition is of dimension $(n+n_c)$.

Denoting $J_{y_{C.L}} \triangleq \lim_{t\to\infty} \frac{1}{t} \int_0^t [y(\bar{p}) + y]^T Q[y(\bar{p}) + y]dt \qquad (4.$

$$J_{y_{n.C.L}} \triangleq \lim_{t\to\infty} \frac{1}{t} \int_0^t [y(\bar{p})^T Q\, y(\bar{p})]dt \qquad (4.$$

$$J_{y_{nICL}} \triangleq \lim_{t\to\infty} \frac{1}{t} \int_0^t [y(\bar{p})^T y(\bar{p})]dt \qquad (4.$$

where the c.l. loop system is given by (4.1), we can compute the above costs as follows:

$$J_{y_{C.L.}} = \text{trace } P_f C_f^T Q_{fo} C_f \quad , \quad Q_{fo} = \begin{bmatrix} Q & Q & 0 & 0 \\ Q & Q & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad (4.$$

$$J_{y_{n.C.L.}} = \text{trace } P_f C_f^T Q_{fn} C_f \quad , \quad Q_{fn} = \begin{bmatrix} Q & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad (4.$$

$$J_{y_{nICL}} = \text{trace } P_f C_f^T \; Q_{fI} \; C_f \quad , \quad Q_{fI} = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.10}$$

The computation of the corresponding quantities for the control variable u is straightforward.

4.2 Measure of Performance Robustness:

Just as in the case of stability robustness, it is helpful even in the case of performance robustness to have a 'measure of performance robustness'. In the case of stability robustness the index $\beta_{S.R}$ represents the degree of nonsingularity of the closed loop system matrix. In similar lines what should the 'Performance Robustness Index' reflect? It may be viewed from many perspectives. In some regulator problems, the maximum allowable r.m.s. output regulation cost derived 'acceptable performance' may be specified, denoted by $J_{y_{n.m}}^{1/2} \triangleq [\int y^T(\bar{p})\; y(\bar{p})]^{1/2}$. The controller designed is then supposed to achieve at least this much output regulation cost, if not lesser. In that case one may view the extent to which a given controller met this requiscent as a measure of performance robustness. In other we can define the 'Performance Robustness Index' as

$$\beta_{P.R.} \triangleq \frac{J_{y_{n.m}} - J_{y_{nICL}}}{J_{y_{n.m}}} \tag{4.11}$$

where $J_{y_{n.m}} \geq J_{y_{nICL}}$, clearly the range of $\beta_{P.R.}$ is given by $0 \leq \beta_{P.R.} \leq 1$. Thus of all the control designs which satisfy the condition $J_{y_{n.m}} \geq J_{y_{nICL}}$, the design with one extreme $\beta_{P.R.} = 1$ has the 'maximum performance robustness' and the other extreme $\beta_{P.R.} = 0$ has zero 'performance robustness'.

In other case, where the cost $J_{y_{n.m}}$ is not specified, one can take the open loop output regulation cost of the evaluation model (i.e., the output regulation cost when there is no control ($u \equiv 0$)) as the maximum value of the output regulation cost. This value can be computed as follows:

$$J_{y_{n.O.L}} = \lim_{t\to\infty} \frac{1}{t} \int_0^t y^T(\bar{p})\, Q\, y(\bar{p})dt \qquad (4.1$$

where the system dynamics is given by the 'evaluation model' equations (3.1), repeated here for convenience,

$$\dot{x} = A\,x + D\,w \quad , \qquad (4.1$$

$$y = C\,x$$

Thus, assuming A is stable,

$$J_{y_{n.O.L}} = \text{trace } P_o C^T Q\, c \qquad (4.1$$

where

$$P_o A^T + A\,P_o + D\,W\,D^T = 0 \qquad (4.1$$

For this case the 'Performance Robustness Index' is

$$\beta_{P.R.} \triangleq \frac{J_{y_{n.C.L}} - J_{y_{n.C.L}}}{J_{y_{n.O.L}}} \qquad (4.1$$

which has the same properties as the $\beta_{P.R.}$ defined by (4.11).

It may be noted that the comments about the maximization of $\beta_{S.R.}$ also holds good for the case of $\beta_{P.R.}$.

Having defined the measures of stability robustness and performance robustness, we are now in a position to present an algorithm to design robust controllers for LQG regulators. The algorithm is elaborated in the next section.

V. An Algorithm to Design Robust Controllers for LQG Regulators:

In this section, based on the results of previous sections, we present an algorithm to design controllers which are robust both from stability as well as performance points of view. For given perturbations in system parameters and model/controller order, the algorithm involves determining the indices $\beta_{S.R.}$ and $\beta_{P.R.}$ for different values of the design parameters ρ_c and ρ_e which in turn determine the control and estimator gains. This information about the indices $\beta_{S.R.}$ and $\beta_{P.R.}$, along with the corresponding values of the costs $(y^T y)^{1/2}$ and $(u^T u)^{1/2}$ can be used to pick a specific controller (control and estimator gain combination) as the one which provides a satisfactory trade-off between stability and performance. The algorithm is thus iterative in nature. The computation basically involves the use of matrix Riccati and Liapunov solutions and the singular value decomposition for which standard, easy-to-use computer programs are available. The algorithm is now summarized in the form of a series of steps to follow. The algorithm presumes the most general case, namely Case 7 of Section (3.3).

5.1 Algorithm to Design a Robust Reduced Order Controller:

Step 1: Data: Read

A, B, C, D, M, n, n_R, n_c, Q, R_o, W, V_o

ΔA, ΔB, ΔC, ΔD, ΔM

Step 2: Store $J_{yn.m}$ of (4.11). If it is not available, assuming A is stable, solve the Liapunov equation (of dimension n)

$$P A^T + AP + DWD^T = 0$$

and compute $J_{yn.O.L} = \text{trace } P C^T Q C$

Step 3: Given n, n_R, obtain the matrices

$$\bar{A}_R, \bar{B}_R, \bar{C}_R, \bar{M}_R, \bar{D}_R$$

where $(\bar{A}_R, \bar{B}_R)$, $(\bar{A}_R, \bar{D}_R)$ c. controllable

and $(\bar{A}_R, \bar{C}_R)$, $(\bar{A}_R, \bar{M}_R)$ c. observable & A_T is asympt. stable

Step 4: Specify ρ_c and ρ_e

Step 5: Solve the Riccati equations

$$K_R\bar{A}_R + \bar{A}_R^T K_R - K_R\bar{B}_R\frac{R_o^{-1}}{\rho_c}\bar{B}_R^T K_R + \bar{C}_R^T Q \bar{C}_R = 0$$

$$P_R\bar{A}_R^T + \bar{A}_R P_R - P_R\bar{M}_R^T\frac{V_o^{-1}}{\rho_e}\bar{M}_R P_R + \bar{D}_R W \bar{D}_R^T = 0$$

and compute $\bar{G}_R = -\frac{1}{\rho_c} R_o^{-1} B_R^T K_R$

$$\bar{\bar{G}}_R = +\frac{1}{\rho_e} P_R M_R^T V_o^{-1}$$

and $$\bar{\bar{A}}_R = \bar{A}_R + \bar{B}_R\bar{G}_R - \bar{\bar{G}}_R\bar{M}_R$$

Step 6: Given n_c, form the matrices

$$\bar{\bar{A}}_{R_{11}}, \bar{\bar{A}}_{R_{12}}, \bar{\bar{A}}_{R_{21}}, \bar{\bar{A}}_{R_{22}}, \bar{\bar{G}}_{R_1}, \bar{\bar{G}}_{R_2}, \bar{G}_{R_1}, \bar{G}_{R_2}$$

Step 7: Given ΔA(i.e. ΔA_R, ΔA_{RT}, ΔA_{TR}, ΔA_T), $\Delta B(\Delta B_R, \Delta B_T)$ $\Delta M(\Delta M_R, \Delta M_T)$, ΔC and ΔD

Form the matrices

$$A_{SC} \text{ and } E_{PPV} \text{ \& } E_{PL}$$

Step 8: Check if $\underline{\sigma}[A_{SC}] > \overline{\sigma}[E_{PL}]$

and $\underline{\sigma}[A_{SC}] > \overline{\sigma}[E_{PPV}]$

If not go to step 4 and repeat the steps 4, 5, 6, 7, 8. If yes, go to step 9.

Step 9: Calculate $\beta_{S.R}$ and store it along with the corresponding values of ρ_c and ρ_e.

Step 10: Form A_f, D_f, C_f, Q_f and W_f

Step 11: Solve the Liapunov equation

$$P_f A_f^T + A_f P_f + D_f W_f D_f^T = 0$$

and calculate $J_{C.L.} = \text{trace } P_f C_f^T Q_f C_f$

$$J_{y.C.L} = \text{trace } P_f C_f^T Q_{fo} C_f$$

$$J_{u.C.L.} = J_{C.L.} - J_{yC.L.}$$

$$J_{y_{n.C.L.}} = \text{trace } P_f C_f^T Q_{f_n} C_f$$

$$J_{y_{nIC.L}} = \text{trace } P_f C_f^T Q_{f_2} C_f$$

and the corresponding quantities with control u.

Step 12: If $J_{y.n.m}$ is given, check if $J_{y.n.m} > J_{y.nICL}$

If $J_{y.n.m}$ is not given, check if $J_{yn.O.L} > J_{yn.C.L}$

If not satisfied, go to step 4 and repeat the steps 4 through 12. If yes go to step 13.

Step 13: Calculate $\beta_{P.R}$

Step 14: Repeat steps 4 through 13 for different values of ρ_c and ρ_e.

Step 15: Plot i) $\beta_{S.R}$ vs. $\beta_{P.R.}$ with the pair (ρ_c, ρ_e) as the parameter ii) $(y^T y)^{1/2}$ vs. $(\sqrt{u^T u})^{1/2}$ with the pair (ρ_c, ρ_e) as the parameter.

Step 16: Select that pair (ρ_c, ρ_e) (and the corresponding controller) that gives acceptable compromize between $\beta_{S.R}$, $\beta_{P.R.}$ $(y^T y)^{1/2}$ and the control effort $(u^T u)^{1/2}$.

5.2 Illustrative Example:

In order to make calculations simple and the methodology clear, we choose a simple first order example to illustrate the design algorithm. The perturbation in this example is taken to be parameter errors alone (with no model/controller truncation).

Example:

Consider a scalar deterministic linear system described by

$$\dot{x} = a\,x + bu, \qquad x(0) = x_o = 1$$

$$y = x \tag{5.}$$

$$z = x$$

with the performance index

$$J = \int_0^\infty (qx^2 + \rho_c\, u^2)dt \tag{5.}$$

where the nominal values of $\bar{a}$ and $\bar{b}$ are a = -1, b = 1 and the state weighing q = 1.

Step 1: $\bar{a} = -1$, $\bar{b} = 1$, $x_o = 1$, $n = 1$, $q = 1$

Let $\Delta a = 0.5$, $\Delta b = 1.207$ (e.g. Δa and Δb could represent the maximum absolute deviation in the direction from $\bar{a}$ and $\bar{b}$).

Step 2: No $J_{y_{n.m}}$ is given. Hence we will compute $J_{y_{n.O.L}}$

Since $\bar{a}$ is stable, $J_{y_{n.O.L}} = +q\,p$

where $2\,p\,\bar{a} + x_o^2 = 0$

$$J_{y_{n.O.L}} = \frac{-q\,x_o^2}{2\bar{a}} = \frac{-1}{-2} = 0.5$$

Step 3: Not necessary since no model/controller reduction is involved.

Step 4: Let $\rho_c = \rho$ (no estimator is needed).

Step 5: Riccati solution:

$$2k\,\bar{a} - \frac{k^2\bar{b}^2}{p} + q = 0$$

$$k = \frac{p}{z\bar{b}^2}\left[2\bar{a} \pm \left(4\bar{a}^2 + \frac{4\bar{b}^2 q}{p}\right)^{1/2}\right]$$

$$g = -\frac{1}{2\bar{b}}\left[2\bar{a} \pm \left(4\bar{a}^2 + \frac{4\bar{b}^2 q}{p}\right)^{1/2}\right]$$

$$a_{C.L} = \bar{a} + \bar{b}\,g$$

For $\rho = \frac{1}{8}$, $g = -2$, $a_{C.L} = -3$

Step 6: Not needed.

Step 7: $a_{C.L} = \bar{a} + \bar{b}\,g = -3$

$$E_{P.V} = \Delta a + \Delta b g = [0.5 + (1.207)(-2)] = -1.9142$$

Step 8: $\underline{\sigma}\ a_{C.L} = 3$

$\overline{\sigma}\ E_{P.V} = 1.9142$

Thus $\underline{\sigma}\ [a_{C.L}] > \overline{\sigma}\ [E_{P.V}]$

Step 9: $\beta_{S.R} = \dfrac{\underline{\sigma}[a_{C.L}] - \overline{\sigma}\ [E_{P.V}]}{\underline{\sigma}[a_{C.L}]} = \dfrac{3 - 1.9142}{3} = 0.3619$

and $\rho = 1/8$

Step 10: $A_f = \begin{bmatrix} a_{C.L} & 0 \\ \Delta a + \Delta bg & a_{C.L} + (\Delta a + \Delta bg) \end{bmatrix}$; $D_f W D_f^T \rightarrow \begin{bmatrix} x_o^2 & 0 \\ 0 & 0 \end{bmatrix}$

Step 11: $J_{y_{n.C.L}} = \dfrac{-qx_o^2}{2\,a_{C.L}} = \dfrac{-1}{-6} = 0.167$

Step 12: $J_{y_{n.O.L}} = 0.5$

$J_{y_{n.C.L}} = 0.167$

Thus $J_{y_{n.O.L}} > J_{y_{n.C.L}}$

Step 13: $\beta_{P.R} = \dfrac{J_{y_{n.O.L}} - J_{y_{n.C.L}}}{J_{y_{n.O.L}}} = \dfrac{0.5 - 0.167}{0.5} = 0.666$

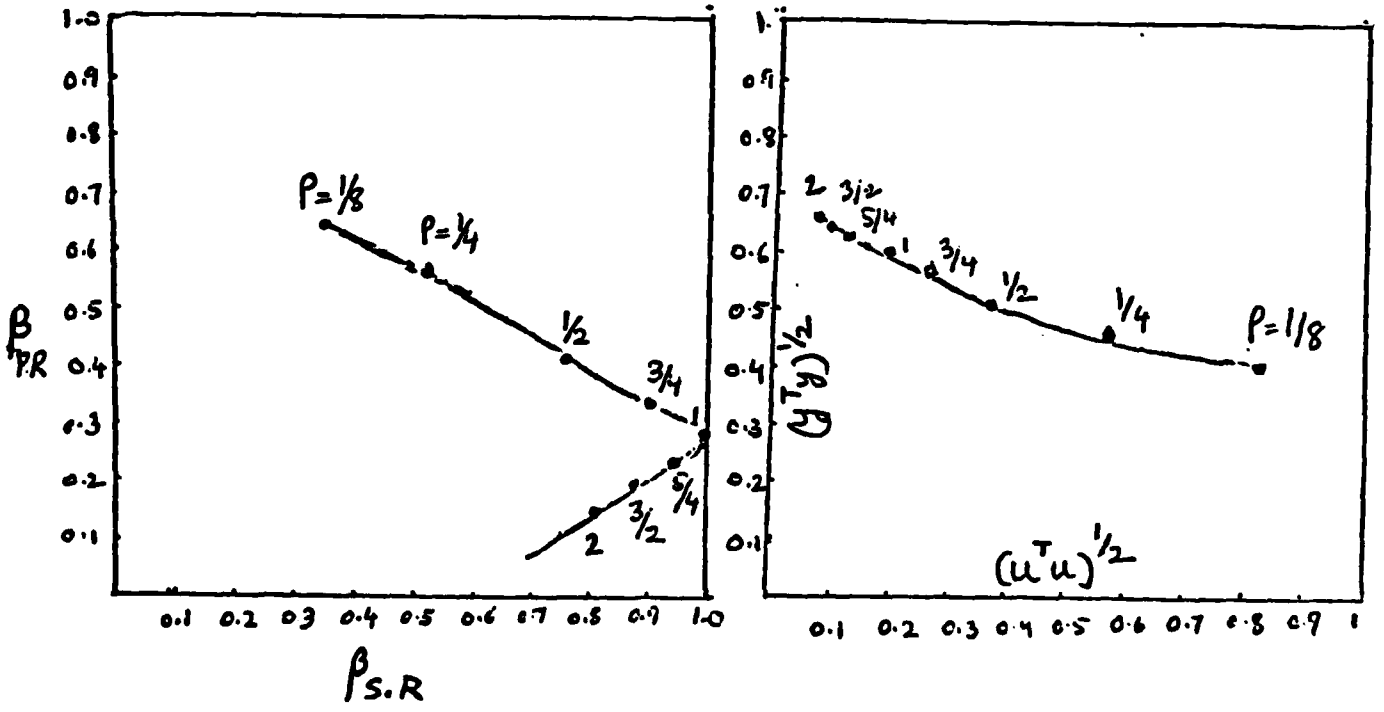
Repeating this procedure for different values of ρ, we obtain the following results.

Step 14:

ρ	$\beta_{S.R}$	$\beta_{P.R}$	$(y^T y)^{1/2}$	$(u^T u)^{1/2}$
1/8	0.362	0.666	0.408	0.817
1/4	0.556	0.553	0.473	0.584
1/2	0.778	0.422	0.537	0.393
3/4	0.911	0.345	0.572	0.301
1	1	0.293	0.595	0.246
5/4	0.935	0.255	0.610	0.208
3/2	0.884	0.225	0.622	0.180
2	0.812	0.183	0.639	0.143

Step 15:

Plots of $\beta_{S.R}$ vs. $\beta_{P.R}$ and $(y^T y)^{1/2}$ vs. $(u^T u)^{1/2}$



Step 16:

From the above plots, one may observe that a reasonable compromize between $\beta_{S.R}$, $\beta_{P.R}$, the regulation cost $(y^T y)^{\frac{1}{2}}$ and the control effort $(u^T u)^{\frac{1}{2}}$ can be achieved with the controller obtained by the weighing $\rho = \frac{1}{4}$. Thus robust controller selected is

$$u = -1.236x$$

VI. Conclusions and Recommendations for Future Research:

The main theme of the summer research has been to incorporate both stability and performance robustness considerations in the design of robust controller for linear regulator problems. First the problem of stability robustness was addressed and conditions for the stability of a perturbed system (where the perturbations are due to parameter errors and mode truncation) were derived using singular value analysis. Next the aspect of performance robustness was addresses and using sensitivity analysis the closed loop system was analyzed for performance. Measures of robustness namely $\beta_{S.R}$ and $\beta_{P.R}$ were defined and based on these a control design algorithm was presented to design robust controllers. The emphasis of the design method has been on the adherence to the standard time domain LQG format. The advantages of this time domain approach are i) tractability of formulation ii) explicit consideration of parameter error and mode truncation information iii) the use of well developed computational techniques such as the solution of Riccati and Liapunov matrix equations and singular valu routines. The limitations are the amount of computation (not complexity) and the perhaps the specification (or generation) of the perturbation matrices ΔA, ΔB, etc. in a given situation.

Portions of the work of this report are being presented at conferences as indicated in refs. [14-16].

As it normally occurs, another result of this study is that many research topics surfaced for further investigation. In the next section, we project those areas which merit serious research effort.

6.2 Avenues for Further Research:

1) The foremost area of research would be to automate the algorithm and apply it to a generic large space structure model. An exercise of this type will be of immense value to understand the intricacies of control design. Incidentally, this particular aspect is being suggested as the topic of investigation under the Minigrant proposal.

2) Another area of interest would be to probe further into the relationship between frequency domain approach and the proposed time domain approach.

3) Extension of the proposed design to discrete time systems could prove useful in the synthesis of digital controllers.

4) The proposed techniques can be used to compare different control strategies from stability and performance points of view.

5) As suggested by Prof. Peter S. Maybeck of the Air Force Institute of Technology, it is of interest to examine the robustness improvement, if any, that can be achieved by considering a wider range of design variables as discussed in Ref. [17].

6) Finally the proposed concepts can be extended to model reduction, sensor actuator placement and critical parameter selection problems.

References

1. Lehtomaki et. al. "Robustness Tests Utilizing the Structure of Modeling Error". Proceedings of Conference on Decision and Control (CDC) 1981, pp. 1173-1190.

2. Lehtomaki, N.A., Sandell, N.R. and Athans, M., "Robustness Results in Linear Quadratic Gaussian based Multivariable Control Designs", IEEE Trans. on Automatic Control, Vol. AC-26, pp. 75-92, 1981.

3. Lee, W.H., Gully S.W., Eterno, J.S. and Sandell, N.R., Jr., "Structural Information in Robustness Analysis", Proceedings of the American Control Conference (ACC), 1982, pp. 1040-1045.

4. Sandell, N.R., Gully, S.W., Lee, W.H. and Lehtomaki, N.A.: "Multivariable Stability Margins for Vehicle Flight Control Systems", Proceedings of CDC, 1981, pp. 1479-1483.

5. Mukhopadhyay, V., and Newton, J.R., "Application of Matrix Singular Value Properties for Evaluating Gain andPhase Margins of Multiloop Systems", Proceedings of the AIAA Guidance and Control Conference, 1982, pp. 420-428.

6. Gupta, N.K., "Frequency Shaped Penalty Functions for Robust Control Design", Proceedings of ACC, 1982, pp. 1035-1039.

7. IEEE Transactions on Automatic Control, Vol. AC-26, No. 1, Feb. 1981.

8. Kosut, R.L., Salzwedel, H., and Emani-Naeini, A,: "Flexible Structure Control: An Analysis of Design Methodologies and Their Robustness Properties", submitted to AIAA Journal of Guidance and Control.

9. Freudenberg, J.S., Looze, D.P. and Cruz, J.B., Jr., "Robustness Analysis using Singular Value Sensitivities", Proceedings of CDC, 1981, pp. 1158-1166.

10. Haddass, Z., and Powell, J.D.: "Design Method for Minimizing Sensitivity to Plant Parameter Variations", AIAA Journal, Vol. 13, Oct. '75, pp. 1295-1303.

11. Ly, V. and Cannon, R.H. ,Jr., "A Direct Method for Designing Robust Optimal Control Systems", Proceedings of AIAA Guidance and Control Conference, Aug. 1978, pp. 440-448.

12. Yedavalli, R.K. and Skelton, R.E., "Controller Design for Parameter Sensitivity Reduction in Linear Regulators" to appear in the Journal of Optimal Control: Applications and Methods, by John Wiley & Sons.

13. Kwakernaak H., and Siran, R., "Linear Optimal Control Systems", Wiley Interscience, 1972.

14. Yedavalli, R.K., "Singular Value Analysis for Stability Robustness of LQG Regulators due to Parameter Variations" to be presented at the AIAA 21st Aerospace Sciences Meeting, Reno, Nevada, Jan. '83.

15. Yedavalli, R.K., "Time Domain Stability Robustness Analysis of LQG Regulators using Singular Value Decomposition" to be presented at the 20th Annual Allerton Conference on Communication, Control and Computing, Univ. of Illinois, Oct. 6-8, 1982.

16. Yedavalli, R.K. "Time Domain Design of Robust Controllers for LQG Regulators", submitted for ACC Conference (1983).

17. Doyle, J.C. and Stein, G., "Robustness with Observers", IEEE Trans. on Automatic Control, AC-24, Aug. 1979, pp. 607-611.

END

FILMED

9-83

DTIC